Foundations of Data Science

Applied Statistics and Probability with Python

Lecture Note

DRAFT

Dr. Md Rezaul Karim
Professor
Department of Statistics and Data Science
Jahangirnagar University, Savar, Bangladesh



'Those who believe, and whose hearts find satisfaction in the remembrance of Allah: for without doubt in the remembrance of Allah do hearts find satisfaction. (Surah Ar-Ra'd (13:28), Al-Quran).

DRAFT

Copyright © 2026 Dr. Md Rezaul Karim

Preface

In today's data-driven world, the ability to analyze and interpret data has become a crucial skill across various disciplines. As we navigate through an abundance of information, the principles of statistics and probability serve as the bedrock upon which data science stands. This book, *Foundations of Data Science: Applied Statistics and Probability with Python*, aims to bridge the gap between theory and practice, providing readers with the tools they need to harness the power of data effectively.

The journey of data science can be both exhilarating and overwhelming. This book is designed for aspiring data scientists, students, and professionals who wish to develop a robust understanding of statistical methods and their application in Python. We will explore key concepts such as descriptive statistics, inferential statistics, hypothesis testing, and probability theory, all while emphasizing practical applications.

Each chapter is structured to introduce fundamental concepts, followed by Python code examples and exercises that reinforce learning. My goal is to make complex ideas accessible and engaging, allowing readers to build their confidence as they apply these techniques to real-world datasets.

I would like to extend my gratitude to the many educators, practitioners, and learners who inspired this work. Your passion for data science fuels my own. I hope this book serves as a valuable resource on your journey, encouraging curiosity and fostering a deeper understanding of the fascinating world of data.

Whether you are a beginner or someone looking to refine your skills, I invite you to dive in and explore the foundations that will empower you in your data science endeavors.

Happy learning!

Prof. Dr. Md Rezaul Karim
June 7, 2026

DRAFT

Table of Contents

1 Introduction to Data Science **1**
- 1.1 Welcome to Data Science . . . 1
- 1.2 Key Components of Data Science . . . 3
- 1.3 Concepts of Statistics . . . 5
  - 1.3.1 Types of Statistics . . . 7
  - 1.3.2 Population and Sample . . . 8
  - 1.3.3 Census and Sample Survey . . . 9
  - 1.3.4 Parameters and Statistic . . . 10
- 1.4 What is Data? . . . 11
- 1.5 Levels of Measurement . . . 13
- 1.6 Variables . . . 16
- 1.7 Scope of Applied Statistics . . . 17
- 1.8 Statistical Methods in Data Science . . . 18
- 1.9 Overview of Data Science Workflow . . . 20
- 1.10 Popular Statistical Analysis Tools . . . 20
- 1.11 Why Choose Python for This Book? . . . 21
- 1.12 Getting Started with Python . . . 23
- 1.13 Structure of the Book . . . 23
- 1.14 Concluding Remarks . . . 23
- 1.15 Chapter Exercises . . . 24

2 Data Summarization and Visualization **26**
- 2.1 Introduction . . . 26
- 2.2 Summarizing Qualitative Data . . . 27
  - 2.2.1 Bar Chart . . . 29
  - 2.2.2 Python Code: Bar Chart . . . 31
  - 2.2.3 Pie Chart . . . 32
  - 2.2.4 Python Code: Pie Chart . . . 34
- 2.3 Summarizing Quantitative Data . . . 36
  - 2.3.1 Line Diagram . . . 36
  - 2.3.2 Frequency Distribution Table . . . 40
  - 2.3.3 Open-Ended Classes in Frequency Distributions . . . 43
  - 2.3.4 Python Code: Frequency Distribution . . . 45

DRAFT

2.3.5 Frequency Polygon 46
2.3.6 Python Code: Frequency Polygon 48
2.3.7 Ogive Curve 49
2.3.8 Python Code: Ogive 51
2.3.9 Histogram 52
2.3.10 Python Code: Histogram 55
2.3.11 Stem-and-Leaf Plot 55
2.3.12 Python Code: Stem-and-leaf 59
2.4 Concluding Remarks 61
2.5 Chapter Exercises 61

3 Data Exploration: Numerical Measures 65
3.1 Introduction 65
3.2 Measures of Central Tendency 66
3.2.1 Arithmetic Mean 66
3.2.2 Advantages and Disadvantages of Arithmetic Mean 68
3.2.3 Geometric Mean 69
Advantages and Disadvantages of Geometric Mean 70
3.2.4 Harmonic Mean 70
Advantages and Disadvantages of Harmonic Mean 71
3.2.5 Relationships Between Arithmetic Mean, Geometric Mean, and Harmonic Mean 72
3.2.6 Median 75
3.2.7 Advantages and Disadvantages of Median 77
3.2.8 Mode 77
3.2.9 Advantages and Disadvantages of Mode 79
3.2.10 Choosing the Ideal Measure of Central Tendency 79
3.2.11 Weighted Mean 80
3.2.12 Measures of Central Tendency for Grouped Data 82
3.2.13 Python Code: Mean, Median and Mode 85
3.2.14 Exercises 86
3.3 Measures of Dispersion or Variability 88
3.3.1 Range 89
3.3.2 Variance 89
3.3.3 Standard Deviation 90
3.3.4 Mean Absolute Deviation (MAD) 92
3.3.5 Measures of Variability for Grouped Data 94
3.3.6 Coefficient of Variation 96
3.3.7 Exercises 97
3.4 Measures of Distribution Shape 98
3.4.1 Skewness 99
3.4.2 Kurtosis 102
3.4.3 Exercises 104
3.5 Measures of Position 105
3.5.1 Quartiles 106

DRAFT

3.5.2 Deciles . . . 109
3.5.3 Percentiles . . . 110
3.5.4 Interquartile Range (IQR) . . . 112
3.5.5 Outlier Identification via IQR Rule . . . 113
3.5.6 Python Code: Dispersion Measures . . . 116
3.5.7 Z-Scores (Standard Scores) . . . 117
3.5.8 Five-Number Summary . . . 118
3.5.9 Boxplot . . . 118
3.5.10 Python Code: Boxplot . . . 124
3.5.11 Exercises . . . 125
3.6 Concluding Remarks . . . 127
3.7 Chapter Exercises . . . 127

4 Introduction to Probability 132
4.1 Introduction . . . 132
4.2 Basic Concepts . . . 133
4.2.1 Experiment . . . 133
4.2.2 Random Experiment . . . 133
4.2.3 Sample Space . . . 134
4.2.4 Events . . . 135
4.3 Probability . . . 135
4.3.1 Types of Probability . . . 136
4.3.2 Properties of the Probability . . . 139
4.3.3 Union of Events . . . 141
4.3.4 Intersection of Events . . . 141
4.3.5 Complementary Event . . . 142
4.3.6 Odds and Odds Ratio . . . 142
4.3.7 Equally Likely Events . . . 145
4.3.8 Mutually Exclusive Events . . . 145
4.3.9 Probability Axioms . . . 146
4.4 Joint and Marginal Probabilities . . . 148
4.4.1 Joint Probability of Events . . . 149
4.4.2 Marginal Probability of Events . . . 149
4.5 Conditional Probability . . . 151
4.6 Independent Events . . . 154
4.7 Posterior Probabilities . . . 159
4.7.1 Law of Total Probability . . . 159
4.7.2 Total Probability with Multiple Conditions . . . 161
4.7.3 Bayes' Theorem . . . 165
4.8 Concluding Remarks . . . 169
4.9 Chapter Exercises . . . 169

DRAFT

5 Random Variable and Its Properties **176**
5.1 Introduction . . . 176
5.2 Random Variable . . . 176
5.3 Discrete Random Variables . . . 178
5.3.1 Probability Mass Function (PMF) . . . 179
5.3.2 Cumulative Distribution Function (CDF) . . . 180
5.3.3 Properties of the Cumulative Distribution Function . . . 181
5.3.4 Exercise . . . 186
5.4 Continuous Random Variables . . . 188
5.4.1 Probability Density Function (PDF) . . . 188
5.4.2 Cumulative Distribution Function . . . 193
5.4.3 Exercises . . . 198
5.5 The Expectation of a Random Variable . . . 199
5.5.1 Example: Testing Electronic Components . . . 200
5.5.2 Example: Metal Cylinder Production . . . 202
5.5.3 Exercises . . . 202
5.6 The Variance of a Random Variable . . . 204
5.6.1 Example: Metal Cylinder Production . . . 207
5.6.2 Chebyshev's Inequality . . . 208
5.6.3 Quantiles of Random Variables . . . 213
5.6.4 Linear Functions of a Random Variable . . . 215
5.6.5 Exercises . . . 219
5.7 Essential Generating Functions . . . 222
5.7.1 Moment Generating Function . . . 223
5.7.2 Key Properties of MGF . . . 223
5.7.3 Probability Generating Function (PGF) . . . 226
5.7.4 Characteristic Function (CF) . . . 230
5.7.5 Key Properties of Characteristic Functions . . . 230
5.7.6 Exercises . . . 236
5.8 Jointly Distributed Random Variables . . . 237
5.8.1 Joint Probability Mass Function (PMF) . . . 237
5.8.2 Example: Daily Customers and Product Sales . . . 238
Verification of Joint PMF . . . 239
Joint Cumulative Distribution Function . . . 239
5.8.3 Joint Probability Density Function . . . 240
5.8.4 Example: Mineral Deposits . . . 240
5.8.5 Marginal Distributions . . . 241
5.8.6 Example: Daily Customers and Product Sales . . . 241
5.8.7 Example: Daily Customers and Product Sales - Conditional Distributions . . . 244
5.8.8 Example: Mineral Deposits . . . 249
5.8.9 Independent Random Variables . . . 250
5.8.10 Covariance and Correlation . . . 252
5.8.11 Linear Combinations of Random Variables . . . 258
5.8.12 Exercises . . . 262

DRAFT

5.9 Python Functions for Statistical Distributions 265
5.10 Concluding Remarks 266
5.11 Chapter Exercises 267

6 Some Discrete Probability Distributions 269
6.1 Introduction 269
6.2 Bernoulli Distribution 269
6.2.1 Expected Value 270
6.2.2 Variance 270
6.2.3 Moment Generating Function 273
6.2.4 Characteristic Function 274
6.2.5 Probability Generating Function 274
6.2.6 Applications 275
6.2.7 Python Code for Bernoulli Distribution 276
6.2.8 Exercises 277
6.3 Binomial Distribution 278
6.3.1 Expected Value 281
6.3.2 Variance and Standard Deviation 282
6.3.3 Example 283
6.3.4 Python Code for Binomial Distribution 287
6.3.5 Exercises 288
6.4 Poisson Distribution 290
6.4.1 Conditions for the Poisson Distribution 290
6.4.2 Expected Value 294
6.4.3 Variance 295
6.4.4 Moment Generating Function 296
6.4.5 Characteristic Function 296
6.4.6 Approximation of Binomial Distribution Using Poisson Distribution 296
6.4.7 Python Code for Poisson Distribution 299
6.4.8 Exercises 300
6.5 The Geometric and Negative Binomial Distributions 303
6.5.1 The Geometric Distribution 304
6.5.2 Expected Value and Variance of Geometric Distribution . 304
6.5.3 Applications of the Geometric Distribution 308
6.5.4 The Negative Binomial Distribution 309
6.5.5 Why Is It Called the Negative Binomial Distribution? . . 313
6.5.6 Relation to Geometric 315
6.5.7 Relationship Between Binomial and Negative Binomial Distributions 316
6.5.8 Probability Generating Function 317
6.5.9 Moment Generating Function 317
6.5.10 Mean and Variance of the Negative Binomial Distribution 320
6.5.11 When to Use the Negative Binomial Distribution 321
6.5.12 Exercises 321

DRAFT

6.6 Discrete Uniform Distribution . . . 323
6.6.1 Mean and Variance of Discrete Uniform Distribution . . . 324
6.6.2 Python Code Discrete Uniform Distribution . . . 328
6.6.3 Exercises . . . 329
6.7 Concluding Remarks . . . 329
6.8 Chapter Exercises . . . 330

7 Some Continuous Probability Distributions 332
7.1 Introduction . . . 332
7.2 Continuous Uniform Distribution . . . 333
7.2.1 Distributional Properties . . . 334
7.2.2 Python Code for Uniform Distribution Characteristics . . . 337
7.2.3 Exercises . . . 338
7.3 Exponential Distribution . . . 339
7.3.1 Memoryless Property . . . 346
7.3.2 Properties of the Exponential Distribution . . . 348
7.3.3 Python Code for Exponential Distribution Characteristics 350
7.3.4 Exercises . . . 352
7.4 Normal Distribution . . . 354
7.4.1 Definition of the Normal Distribution . . . 355
7.4.2 Properties of the Normal Distribution . . . 358
7.4.3 Inflection Points of the Normal Distribution . . . 360
7.4.4 Standard Normal Distribution . . . 362
7.4.5 Relationship between X and Z (Z-score) . . . 365
7.4.6 Central Limit Theorem . . . 375
7.4.7 Approximating the Binomial Distribution with the Normal Distribution . . . 379
7.4.8 Python Code for Normal Distribution Characteristics . . . 382
7.4.9 Exercises . . . 384
7.5 Concluding Remarks . . . 386
7.6 Chapter Exercises . . . 387

8 Confidence Interval Estimation 389
8.1 Introduction . . . 389
8.2 Confidence Intervals . . . 390
8.3 Confidence Intervals for the Population Mean . . . 390
8.4 Confidence Intervals for Variances and Standard Deviations . . . 405
8.4.1 Confidence Interval for Variance . . . 405
8.4.2 Confidence Interval for Standard Deviation . . . 406
8.5 Confidence Intervals for Population Proportions . . . 411
8.6 Sample Size Estimation . . . 414
8.6.1 Sample Size for Estimating a Population Mean . . . 414
8.6.2 Sample Size for Estimating a Population Proportion . . . 418
8.6.3 Sample Size Estimation for Finite Populations . . . 421
8.7 Concluding Remarks . . . 423

DRAFT

8.8 Chapter Exercises 423

9 Hypothesis Testing for Decision Making 427
9.1 Introduction 427
9.2 Concepts of Hypothesis Testing 428
9.3 Steps for Hypothesis Testing 429
9.3.1 Formulating Hypotheses 430
9.3.2 Level of Significance 431
9.3.3 Test Statistics 433
9.3.4 Acceptance and Rejection Regions 433
9.3.5 Decision Rules 435
9.4 The p-value 436
9.5 Hypothesis Testing for Means 437
9.5.1 One-Sample Test of Means 437
9.5.2 Testing Equality of Two Means 449
9.5.3 Independent Samples T-Test 449
1. Equal Variances (Pooled T-Test) 450
2. Unequal Variances (Welch's T-Test) 451
9.5.4 Paired T-Test 452
9.6 Testing Equality of Several Means 455
9.6.1 Analysis of Variance (ANOVA) 455
9.7 Tests for Variance 464
9.7.1 One-Sample Variance Test 464
9.7.2 Two-Sample Variance Test (F-Test) 466
9.7.3 Confidence Intervals for Variance 470
Confidence Interval for a Single Variance 470
Confidence Interval for the Ratio of Two Variances 471
9.7.4 Tests for Homogeneity of Variance Across Multiple Groups 471
Bartlett's Test 471
Levene's Test 472
Comparison of Bartlett's and Levene's Tests 473
Post-Hoc Analysis 474
9.7.5 Practical Considerations 475
9.8 Power of the Test 476
9.9 Sample Size Estimation for the Mean Test 478
9.9.1 When Testing for the Mean of a Normal Distribution (One-Sided Alternative) 479
9.9.2 Sample Size Estimation When Testing for the Mean of a Normal Distribution (Two-Sided Alternative) 480
9.10 Sample Size Estimation for Proportion Test 483
9.11 Concluding Remarks 486
9.12 Chapter Exercises 486

DRAFT

10 Nonparametric Tests of Hypothesis **492**
10.1 Introduction . . . 492
10.2 Characteristics of Nonparametric Tests . . . 493
10.3 When to Use Nonparametric Tests . . . 495
10.4 The Sign Test . . . 498
10.4.1 Test Statistic . . . 498
10.4.2 Hypotheses and Decision Rule . . . 498
10.4.3 Wilcoxon Signed-Rank Test . . . 501
Hypotheses . . . 501
Test Procedure . . . 502
Large Sample Approximation . . . 502
10.5 The Mann–Whitney U Test . . . 505
10.6 The Kruskal–Wallis Test . . . 513
10.7 The Friedman Test . . . 516
10.7.1 The Friedman Test vs. The Kruskal–Wallis Test . . . 518
10.8 Chi-Square Test for Goodness-of-Fit . . . 519
10.9 The Pearson Chi-Square Test for Categorical Data . . . 523
10.10 Yates's Continuity Correction for 2×2 Contingency Tables . . . 527
10.11 Fisher's Exact Test for 2×2 Tables . . . 530
10.12 Kolmogorov–Smirnov Test for Goodness-of-Fit . . . 538
10.12.1 One-Sample K–S Test . . . 539
10.12.2 Two-Sample K–S Test . . . 541
10.12.3 Advantages and Limitations . . . 542
10.12.4 Implementation in Software . . . 542
10.13 Chapter Exercises . . . 542

11 Correlation and Regression Analysis **546**
11.1 Introduction . . . 546
11.2 Scatter Diagram . . . 547
11.3 Python Code: Scatter diagram . . . 549
11.4 Correlation Analysis . . . 550
11.5 Covariance . . . 550
11.6 Pearson's Correlation Coefficient . . . 555
11.6.1 Interpretation of the value of Correlation Coefficient . . . 556
11.6.2 Properties of the Correlation Coefficient . . . 557
11.6.3 Testing the Significance of the Correlation Coefficient . . . 561
11.6.4 Python Code: Correlation Matrix . . . 562
11.7 Rank Correlation . . . 563
11.7.1 Key Types of Rank Correlation . . . 563
11.7.2 Applications of Rank Correlation . . . 566
11.7.3 Python Code: Rank Correlation . . . 566
11.7.4 Kendall Tau Correlation Coefficient . . . 567
11.7.5 Advantages and Disadvantages . . . 570
11.7.6 Python Code: Kendall Tau . . . 570
11.7.7 Exercises . . . 570

DRAFT

11.8 Regression Analysis . . . 574
11.8.1 Types of regression analysis . . . 576
11.8.2 Simple Regression Model . . . 577
11.8.3 Assumptions of the CSLR Model (11.6) . . . 579
11.8.4 Ordinary Least Squares (OLS) Estimation . . . 581
11.8.5 Interpretation of Regression Coefficients . . . 583
11.8.6 The Estimated Error Variance or Standard Error . . . 584
11.8.7 Coefficient of Determination . . . 590
11.8.8 Relationship between R^2 and r . . . 591
11.8.9 Advantages and Disadvantages of R^2 . . . 594
11.8.10 Adjusted R^2 . . . 594
11.8.11 Relationship between R^2 and $\bar{R}^2$ in Simple Linear Regression . . . 595
11.8.12 Interval Estimation and Hypothesis Testing . . . 605
11.8.13 The t-tests in Simple Linear Regression Model . . . 606
11.8.14 Confidence Interval for $E(Y|X = x)$ . . . 607
11.8.15 Python Code: Linear Regression Model . . . 609
11.8.16 Exercises . . . 610
11.9 Multiple Linear Regression . . . 618
11.9.1 Model Assumptions . . . 618
11.9.2 Estimation Procedure . . . 619
11.9.3 Estimation Procedure of Error Variance . . . 620
11.9.4 Mean of the OLS Estimator . . . 620
11.9.5 Variance of the OLS Estimator . . . 621
11.9.6 Coefficient of Determination . . . 621
11.9.7 Adjusted R^2 . . . 622
11.9.8 Example Dataset and Regression Calculations . . . 623
11.9.9 F-test in Multiple Regression . . . 626
11.9.10 ANOVA Table in Regression Analysis . . . 626
11.9.11 The t-tests in Multiple Regression . . . 627
11.9.12 Exercises . . . 636
11.10 Regression Model Diagnostics . . . 637
11.10.1 Assumptions of Linear Regression . . . 638
11.10.2 Residual Plots . . . 638
11.10.3 Formal Tests . . . 643
11.10.4 Test for Autocorrelation . . . 646
11.10.5 Tests for Non-Constancy of Variance . . . 647
11.10.6 Influential Observations, Outliers, and Cook's Distance . . . 653
11.10.7 Multicollinearity . . . 655
11.10.8 Exercises . . . 660
11.11 Concluding Remarks . . . 661
11.12 Chapter Exercises . . . 661

DRAFT

12 Applied Time Series Analysis and Forecasting 672
12.1 Introduction . . . 672
12.2 What is a Time Series? . . . 673
12.2.1 Time series Plot . . . 675
12.2.2 Example: AirPassengers dataset . . . 675
12.2.3 Importance of Time Series Analysis . . . 677
12.2.4 Time Series Data vs. Cross-Sectional Data . . . 677
12.2.5 Pooled Data vs. Longitudinal Data . . . 677
12.2.6 Types of Time Series . . . 678
12.2.7 Time Series Forecasting . . . 678
12.2.8 Preliminary Adjustments Before Analyzing Time Series . 679
12.2.9 Components of a Time Series . . . 680
12.2.10 Methods of Time Series Analysis . . . 681
12.3 Main Components of a Time Series . . . 683
12.4 Mathematical Models for Time Series . . . 684
12.4.1 Additive Model . . . 684
12.4.2 Multiplicative Model . . . 684
12.4.3 Mixed Models . . . 685
12.5 Classical Decomposition . . . 686
12.6 STL Decomposition . . . 688
12.7 Measurement of Trend and Forecasting . . . 692
12.7.1 Freehand or Graphic Method . . . 692
12.7.2 Method of Semi-Averages . . . 693
12.7.3 Method of Moving Averages . . . 695
12.7.4 Method of Fitting Curves . . . 704
12.8 Measurement of Seasonal Variation . . . 712
12.8.1 Method of Simple Averages . . . 713
12.8.2 Ratio to Trend Method . . . 717
12.8.3 Ratio-to-Moving-Average Method . . . 725
12.8.4 Method of Link Relatives . . . 730
12.9 Measuring Cyclic Variation . . . 737
12.9.1 Residual Method . . . 737
12.10 Measuring Irregular Component . . . 739
12.11 Key Statistical Properties of Time Series . . . 740
12.11.1 Mean Function . . . 740
12.11.2 Variance Function . . . 740
12.11.3 Autocovariance Function . . . 741
12.11.4 Autocorrelation Function . . . 741
12.11.5 Correlogram . . . 742
12.11.6 Partial Autocorrelation Function (PACF) . . . 744
12.11.7 White Noise . . . 748
12.12 Stationary Processes . . . 750
12.13 Random Walk Model . . . 752
12.14 Stationarity Test . . . 756
12.14.1 Dickey-Fuller Test . . . 757

DRAFT

12.14.2 Augmented Dickey-Fuller (ADF) Test 758
12.14.3 Phillips-Perron (PP) Test 761
12.14.4 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test 763
12.15 Models for Stationary Time Series 765
12.15.1 Autoregressive (AR) Model of Order p 765
12.15.2 Example: AR(1) Model 771
12.15.3 Parameter Estimation of AR(p) Model 774
12.15.4 Moving Average (MA) Model 776
12.15.5 Example: MA(1) Model 779
12.15.6 Autoregressive Moving Average (ARMA) Model 780
12.15.7 Example: ARMA(1,1) Model 785
12.15.8 Example: ARMA(1,1) Model 786
12.16 Models for Nonstationary Time Series and Forecasting 788
12.16.1 Stationarity and Differencing 788
12.16.2 ARIMA(p, d, q) Model 789
12.16.3 Model Identification 790
12.16.4 Estimation of Parameters in ARIMA Models 791
12.16.5 Forecasting Equation of ARIMA(p, d, q) Model 793
12.16.6 Model Diagnostics 794
12.17 The Seasonal ARIMA Model 796
12.17.1 The SARIMA Model Equation 797
12.17.2 Example: A SARIMA$(0, 1, 1)(0, 1, 1)_{12}$ Model 797
12.17.3 Example: A SARIMA$(1, 1, 1)(1, 1, 1)_{12}$ Model 798
12.17.4 Parameter Estimation 799
12.17.5 Python Code: SARIMA$(p, d, q)(P, D, Q)_s$ Model 801
12.18 Exponential Smoothing and the ETS Framework 803
12.18.1 Exponential Smoothing 803
12.18.2 Simple Exponential Smoothing 804
12.18.3 Holt's Linear Trend Model 807
12.18.4 Holt-Winters Seasonal Model 809
12.19 The ETS Framework 813
12.19.1 Components of the ETS Framework 813
12.19.2 Common ETS Models 815
12.19.3 Mathematical Formulation 815
12.20 Machine Learning Techniques for Time Series Analysis 817
12.20.1 ANN for Time Series Forecasting 817
12.20.2 Feedforward Neural Networks (FNNs) 818
12.20.3 Recurrent Neural Networks (RNNs) 824
12.21 LSTM for Time Series Forecasting 827
12.21.1 LSTM Architecture for Time Series 827
12.21.2 Practical Example: Stock Price Forecasting 829
12.21.3 Key Advantages of LSTM for Time Series 835
12.21.4 Training Considerations 835

13 Appendix **836**

DRAFT

Chapter 1

Introduction to Data Science

1.1 Welcome to Data Science

Welcome to the exciting world of data science, where numbers tell hidden stories and reveal valuable insights! In today's digital era, data is incredibly powerful, and those who can understand and use it hold the key to endless opportunities. Data science is all about extracting meaningful information from vast amounts of data to make informed decisions, solve complex problems, and drive innovation.

> **Data Science:** A multidisciplinary field focused on the collection, processing, analysis, and interpretation of large volumes of structured and unstructured data using methods from statistics, machine learning, and computer science. Its goal is to uncover patterns, generate insights, and drive evidence-based decision-making across a wide range of industries and research domains.

In the era of **big data** and advanced analytics, data science has become a crucial field in both academia and industry. At the heart of data science is the ability to extract meaningful insights from data, a process that heavily relies on statistical methods. To succeed in this field, a strong foundation in statistics is essential. Statistics provides the tools and techniques needed to analyze data, identify patterns, and draw reliable conclusions. Without this knowledge, it is challenging to make sense of the data and leverage its full potential.

While data science may appear to be a modern phenomenon, its roots trace back several decades. The field evolved gradually through key contributions from statisticians, computer scientists, and industry pioneers.

DRAFT

A Brief History of Data Science

The journey begins in 1962, when renowned statistician **John Tukey** published his visionary essay *The Future of Data Analysis* [19]. He called for a new scientific discipline focused on exploring and interpreting data, going beyond traditional statistical inference. Tukey emphasized empirical analysis and visualization, concepts that are now central to modern data science.

In 1974, Danish computer scientist **Peter Naur** proposed the term *Data Science* as an alternative name for computer science in his book *Concise Survey of Computer Methods* [15]. Though his definition centered on data processing rather than analytics, it marked one of the earliest uses of the term.

The idea that data science could serve as a modern identity for statistics was further advanced by **C.F. Jeff Wu**. In his influential lectures during the 1980s and 1990s, he argued for renaming statistics as "data science" to reflect its evolving role and broader application [21]. His lecture, titled *Statistics = Data Science?*, highlighted the need to break away from outdated stereotypes and embrace a more applied, interdisciplinary focus.

A major milestone came in 2001, when **William S. Cleveland** published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics* [4]. In this foundational paper, Cleveland outlined a comprehensive plan for data science that integrated statistics with computing, visualization, and collaboration across domains. His work provided a practical and academic framework for establishing data science as a formal field of study.

The term **Data Scientist** gained widespread recognition in 2008, when **DJ Patil** (LinkedIn) and **Jeff Hammerbacher** (Facebook) began using it to describe professionals capable of working with massive datasets to generate business insights [16]. Their adoption of the title helped popularize it across the tech industry and beyond.

Today, data science continues to evolve, integrating machine learning, artificial intelligence (AI), and cloud-based data platforms. It plays a critical role in solving real-world problems, whether it is forecasting economic trends, improving healthcare outcomes, or optimizing user experiences in digital applications.

Your Journey Starts Here

In this chapter, we will introduce the foundational concepts of applied statistics that are essential for data scientists. We will cover basic statistical terminology, explore different types of data, and discuss various statistical methods commonly used in data science. But data science is more than just numbers and formulas, it involves programming, data engineering, machine learning, and

DRAFT

data visualization. These skills enable data scientists to collect, clean, and process data, build predictive models, and communicate results effectively.

As we embark on this journey, we will dive deep into the statistical foundations that every aspiring data scientist needs to master. We will also explore how these principles are applied in real-world scenarios, from predicting customer behavior to identifying trends in healthcare. So, buckle up and get ready to unlock the secrets of data science!

1.2 Key Components of Data Science

Data Science is made up of several interrelated components, each playing a crucial role in transforming raw data into actionable insights. Below are some of the most important components:

Statistics and Probability

Statistics and probability form the mathematical foundation of Data Science, providing essential tools for analyzing data, making inferences, and building predictive models. Statistics helps summarize and describe data through measures such as mean, median, variance, and standard deviation, while also enabling hypothesis testing and confidence interval estimation. Probability theory is used to model uncertainty and assess the likelihood of different outcomes, which is especially crucial in decision-making and risk analysis. Together, these disciplines support key Data Science tasks such as data interpretation, model evaluation, and the development of algorithms in machine learning. Tools like R, Python (with libraries such as `NumPy`, `SciPy`, and `Statsmodels`), SPSS and SAS are commonly used for statistical analysis.

Machine Learning

Machine learning, a subset of artificial intelligence, plays a central role in Data Science by enabling computers to learn from data and make predictions or decisions without explicit programming. It involves training models on historical data to detect patterns and generate accurate forecasts or classifications. Machine learning is widely applied in areas such as recommendation systems, fraud detection, image recognition, and natural language processing. Common approaches include supervised learning, unsupervised learning, and reinforcement learning. Popular tools and libraries used to build and deploy models include `Scikit-learn`, `TensorFlow`, `Keras`, and `PyTorch`.

Deep learning

Deep learning is an advanced subset of machine learning that uses artificial neural networks with multiple layers to model and understand complex pat-

DRAFT

terns in large volumes of data. It excels at tasks where traditional algorithms struggle, such as image and speech recognition, natural language processing, and autonomous systems. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), learn hierarchical features directly from raw data without the need for manual feature extraction. These models require significant computational power and large datasets to perform effectively. Popular frameworks for developing deep learning applications include `TensorFlow`, `Keras`, and `PyTorch`.

Data Engineering

Data Engineering involves the design, development, and maintenance of systems and architectures that enable the collection, storage, and processing of large datasets. Data engineers build data pipelines that move data from various sources to storage and analytics platforms, ensuring it is clean, reliable, and accessible for analysis. They work with tools and technologies like SQL, Apache Spark, Hadoop, Airflow, and cloud services (e.g., AWS, Azure, GCP) to handle big data efficiently. Data Engineering lays the foundation for data analysis and machine learning by making high-quality data available to data scientists and analysts.

Data Visualization

Data visualization is a critical aspect of Data Science that involves representing data and analytical results through visual formats such as charts, graphs, maps, and dashboards. It helps transform complex datasets into easily understandable insights, enabling quicker interpretation and more informed decision-making. Effective data visualization allows analysts and stakeholders to identify patterns, trends, and outliers that might not be immediately evident in raw data. It is widely used in business intelligence, reporting, and exploratory data analysis. Common tools and libraries include `Tableau`, `Power BI`, `Matplotlib`, `Seaborn`, and `Plotly`, each offering powerful capabilities for creating both static and interactive visualizations.

Big Data Analytics

Big Data Analytics is a key component of Data Science that focuses on processing and analyzing vast volumes of complex and fast-moving data. It enables data scientists to uncover meaningful patterns, trends, and insights from large datasets that traditional tools cannot handle efficiently. By leveraging technologies like Hadoop, Spark, and NoSQL databases, Big Data Analytics supports real-time decision-making and predictive modeling. This component is essential in industries where data is generated at high speed and scale, helping organizations drive innovation, improve operations, and gain a competitive edge.

DRAFT

In this book, we focus on statistics and probability, emphasizing applied statistical tools and applied probability with examples. The following chapters will provide detailed explanations and practical applications.

1.3 Concepts of Statistics

Understanding the concepts of statistics is crucial for any data scientist. Statistics provides the foundation for data analysis, enabling us to make sense of data and draw meaningful conclusions. As a data scientist, you will frequently encounter questions such as:

- How do we summarize large datasets?
- What can we infer from sample data about a larger population?
- How can we validate the accuracy of our models?
- What techniques can we use to identify patterns or anomalies in data?

By mastering statistical concepts, you will be better equipped to answer these questions and make informed decisions based on data. Statistics plays a crucial role in data science by providing tools and methods for:

- Summarizing and exploring data
- Designing experiments and surveys
- Making inferences and predictions
- Evaluating and validating models

In this section, we will cover some fundamental statistical concepts essential for data science.

Statistics: The science of collecting data, organizing, summarizing, classifying, comparing, and drawing inferences about a population.

Statistics encompasses both theoretical and applied aspects. The theoretical side focuses on developing new statistical methods and theories. Applied statistics involves the practical application of these methods to real-world problems and data, utilizing established statistical techniques to analyze data and draw conclusions in various fields such as healthcare, business, engineering, and social sciences.

Applied statistics forms the foundation of data science by collecting, organizing, and analyzing data to understand larger populations. By applying these methods, experts can predict trends and outcomes using models based

DRAFT

on real-world observations. This process enhances data quality through experiments and sampling techniques and supports evidence-based conclusions. Applied statistics is vital for fields like natural sciences, social sciences, public policy, healthcare, engineering, and business, enabling informed decisions and systematic approaches to complex challenges.

Example: Customer Satisfaction Survey

Imagine a company wants to understand the average customer satisfaction level with its new product. Surveying every customer is impractical, so the company surveys a sample of 500 customers out of its 50,000 customers.

Collecting Data

The company distributes a satisfaction survey to 500 randomly selected customers, asking them to rate their satisfaction on a scale from 1 to 10.

Analyzing Data

Once the responses are collected, the company calculates the average satisfaction score from the sample. Suppose the average satisfaction score from the 500 customers is 7.2.

Interpreting Data

Using statistical methods, the company estimates the average satisfaction score for the entire population of 50,000 customers based on the sample. This involves calculating confidence intervals to understand the range within which the true average satisfaction score likely falls.

Presenting Data

The company creates a report with graphs and charts to visualize the distribution of satisfaction scores, the average satisfaction score, and the confidence interval.

Organizing Data

The company stores the survey data in a database, ensuring it is organized for future analysis or reference.

Making Inferences

By analyzing the sample data, the company infers that the average satisfaction score of all its customers is approximately 7.2, with a certain level of confidence (e.g., 95% confidence interval).

DRAFT

Quantifying Uncertainty

To quantify the uncertainty of their estimate, the company calculates a confidence interval. For instance, they might determine that they are 95% confident that the true average satisfaction score of all customers is between 6.9 and 7.5.

Through collecting, analyzing, interpreting, presenting, and organizing data from a sample, the company uses statistics to make informed decisions about customer satisfaction for the entire customer base. This approach helps the company understand and quantify the uncertainty associated with their estimate, enabling them to make more accurate and reliable business decisions.

1.3.1 Types of Statistics

When using statistics to derive information from data for decision-making, we employ either descriptive statistics or inferential statistics. The choice between these methods depends on the questions we aim to answer and the nature of the data at hand.

(i). Descriptive Statistics

Descriptive statistics is all about **summarizing data** in a way that is easy to understand. It involves methods used to summarize and organize the main features of a dataset, providing a clear and concise overview without making broader generalizations. It focuses on describing the characteristics of the data at hand, employing tools like measures of central tendency (mean, median, mode) and measures of variability (range, variance, standard deviation), along with graphical representations such as histograms and bar charts.

> **Descriptive statistics:** Methods of organizing, summarizing, and presenting data in an informative way.

The primary goal is to present raw data in an understandable format, simply describing what the data shows rather than drawing conclusions about a larger population.

(ii). Inferential Statistics

Inferential statistics allows us to make predictions or generalizations about a larger group based on a small sample of data. In many situations, it's not practical or possible to collect data from everyone (for example, asking every person in a country about their opinion). Instead, we collect data from a small group (a sample) and use that to make guesses about the entire group (called a population).

For example, suppose you wanted to know how much time students spend studying for exams in your school. Instead of asking every student, you could

DRAFT

ask a small group of students (a sample) and then use their answers to make an estimate about the entire student population.

> **Inferential statistics:** Techniques for making predictions or inferences about a population based on sample data.

Inferential statistics helps us **make decisions** or **draw conclusions** about a larger group, even though we only have data from a smaller group. This is helpful when we need to make choices or predictions without needing to survey everyone.

1.3.2 Population and Sample

In statistics, it is often impractical or impossible to study an entire population. Instead, we rely on samples. Understanding the difference between a population and a sample is fundamental to conducting statistical analyses.

Population

The population refers to the entire group of individuals or instances about whom we want to draw conclusions. It includes all possible observations or outcomes that are of interest in a particular study or analysis.

> **Population:** The entire set of individuals, items, or data points of interest in a study or analysis.

Examples:

- If we are studying the prevalence of diabetes in adults aged 40-60 in a country, the population would include all adults aged 40-60 in that country. This would encompass every individual in that age range, regardless of their health status, socioeconomic background, or other characteristics.
- If we are studying the average height of adult men in a country, the population would include all adult men in that country.

Sample

A sample is a subset of the population that is selected for the actual study. The goal is to choose a sample that is representative of the population so that the findings can be used to make inferences or generalizations about the population.

> **Sample:** A subset of a population selected for analysis to draw conclusions about the entire population.

DRAFT

Examples:

- Continuing with the diabetes study, it might be impractical to test every adult aged 40-60 in the country. Instead, researchers might select a sample of 1,000 adults from this age group and measure their blood sugar levels. The results from this sample can then be used to estimate the prevalence of diabetes in the entire population of adults aged 40-60.

- If we cannot measure the height of all adult men in a country, we might select a sample of 1000 adult men and measure their heights. The results from this sample can then be used to estimate the average height of the entire population.

By studying a sample, we can make inferences about the population. However, it is crucial to use proper sampling techniques to ensure that the sample accurately represents the population.

1.3.3 Census and Sample Survey

Census

The census is a comprehensive and periodic data collection process aimed at gathering detailed demographic information about the entire population of a country. In the context of Bangladesh, the census is conducted by the Bangladesh Bureau of Statistics (BBS) and serves as a crucial tool for policy planning, resource allocation, and development programs. The first population census in Bangladesh was conducted in 1974, following the country's independence. Since then, the census has been carried out every ten years, with the most recent one being the **2022 Population and Housing Census**. As per as 2022 census, Bangladesh have a population of 165,158,616 people, of which 81,712,824 are male, while 83,347,206 are female. As many as 113,063,587 of them live in rural areas and 52,009,072 live in Urban.

Importance of Census

The census data is essential for:

- Informing government policies and development strategies.

- Allocating resources and planning public services at both national and local levels.

- Providing a basis for the estimation of various demographic trends and indicators.

DRAFT

Sample Surveys

Sample surveys are a method of collecting data from a subset of the population to infer information about the entire population. In Bangladesh, sample surveys are conducted for various purposes, including economic research, social policy evaluation, and program assessment.

Types of Sample Surveys

In Bangladesh, different types of sample surveys are carried out by the Bangladesh Bureau of Statistics (BBS) and other organizations. Some key surveys include:

- **Labor Force Survey (LFS):** Assesses employment, unemployment, and labor market conditions.
- **Multiple Indicator Cluster Survey (MICS):** Provides data on child health, education, and protection.
- **Household Income and Expenditure Survey (HIES):** Measures household income, consumption patterns, and poverty levels.
- **Demographic and Health Survey (DHS):** Collects data on population health, fertility, and mortality.

1.3.4 Parameters and Statistic

Parameter

A **parameter** is a measurable characteristic of the population which is a numerical value that summarizes a characteristic of a population. It is a fixed value, although its exact value is often unknown. Parameters describe the entire population.

> **Parameter:** A numerical characteristic or constant factor that describes an entire population, such as its mean, variance, or proportion.

Examples:

- **Population Mean (μ):** The average height of all adult men in a country.
- **Population Proportion (P):** The proportion of voters in a country who support a particular candidate.

Statistic

A **statistic** is a measurable characteristic of the sample which is a numerical value that summarizes a characteristic of a sample. It is used to estimate the corresponding population parameter. Statistics can vary from sample to sample.

DRAFT

Statistic: A numerical characteristic calculated from a sample of data, used to estimate a population parameter.

Examples:

- **Sample Mean ($\bar{x}$):** The average height of 100 randomly selected adult men from the population.
- **Sample Proportion (p):** The proportion of voters in a sample of 1,000 who support a particular candidate.

Parameters describe the entire population and are fixed values, whereas statistics describe samples and can vary from sample to sample. Understanding the distinction between these concepts is crucial in the field of statistics, as we often use sample statistics to make inferences about population parameters.

1.4 What is Data?

Data are raw facts, observations, or units of information, including numbers, words, measurements, observations, images, videos, audio, or descriptions, that can be collected, stored, analyzed, and used to inform decisions. They can be structured (like databases) or unstructured (like text or images). They are the basis for generating useful insights and making informed decisions in various fields.

Data: Raw facts and figures or units of information that can be collected for analysis, which can be quantitative or qualitative.

It is important to distinguish between **data** and **information**. Data, in its raw form, may not convey immediate meaning. When data is processed, organized, and given context, it becomes information, which is then valuable for understanding and making decisions

Types of Data

Data can be categorized into several types, each with its own characteristics and uses. These categories help us understand different ways to collect and analyze information. The primary data types are as follows:

(i). **Quantitative Data or Numeric Data**: This type of data involves numerical values that can be measured or counted, answering questions like "how much?" or "how many?" Quantitative data provides us with specific amounts or counts, and it can be broken down further into two subcategories:

DRAFT

- **Discrete Data**: It refers to numerical data that can take on only specific, distinct, and separate values—typically countable quantities. These values usually arise from counting (e.g., number of students, number of cars). Integer values are often used to represent discrete data because they cannot be meaningfully subdivided in the context of the data (e.g., you can not have 10.5 students).
- **Continuous Data**: In contrast, continuous data can take any value within a given range, including fractions. Examples include measurements like height, weight, or temperature.

Some examples of quantitative data are as follows:

- How old you are: 15, 16, 17, etc.
- How many students are in a class: 25, 30, 35, etc.
- The temperature in a city: 20°C, 25°C, etc.

(ii). **Qualitative Data or Non-numeric Data**: Unlike quantitative data, qualitative data or non-numeric data does not deal with numbers. Instead, it describes categories, qualities, or characteristics. This type of data is often used to answer questions like "what kind?" or "which one?" Sometimes, it is called **categorical data**. Qualitative data includes various forms of descriptive information, and examples include:

- Your favorite color: Red, Blue, Green, etc.
- The type of pets people have: Dog, Cat, Fish, etc.
- The kind of food people like: Pizza, Pasta, Salad, etc.

Other Types of Data

- **Binary Data**: Contains only two possible values (e.g., true/false, yes/no).
- **Text Data**: Unstructured information in textual form (e.g., articles, tweets).
- **Time Series Data**: Recorded observations over time intervals (e.g., stock prices, weather data).
- **Spatial Data**: Information about the physical location and shape of objects (e.g., GPS coordinates, maps).
- **Image and Video Data**: Visual information captured as images or videos.

DRAFT

- **Audio Data**: Sound or speech information, including recordings or live streams.

1.5 Levels of Measurement

The **levels of measurement** or **scale of measurement** refer to the nature of the data and determine the types of statistical analyses that can be performed. There are four main levels of measurement:

(i). Nominal

(ii). Ordinal

(iii). Interval

(iv). Ratio

Nominal Level

The nominal level of measurement is the most basic type of data categorization. It classifies data into distinct categories that do not have a natural order or ranking. These categories are mutually exclusive and collectively exhaustive, meaning each observation fits into one and only one category, and all categories together include all possible observations.

Characteristics of Nominal Data:

- Categories are distinct and mutually exclusive.
- No inherent order or ranking among categories.

Examples:

- Gender (male, female)
- Marital status (single, married, divorced)
- Types of cuisine (Italian, Chinese, Mexican)
- Types of Engineering Degrees: Civil, Mechanical, Electrical, Chemical, Computer.
- Different types of engines used in vehicles and machinery: Diesel, Electric, Gasoline, Hybrid.
- Various materials used in building structures: Wood, Steel, Concrete, Brick.
- Blood Types: A+, A-, B+, B-, AB+, AB-, O+, and O-.

DRAFT

- Disease Types: Flu, Cold, Allergies, Asthma.
- Medical Specialties: Cardiology, Neurology, Oncology, Pediatrics.
- Vaccination Status: Vaccinated, Not Vaccinated.
- Hospital Departments: Emergency, Radiology, Surgery, Pediatrics.

Ordinal Level

The ordinal level of measurement classifies data into categories that have a meaningful order or ranking among them. Unlike nominal data, ordinal data allow for comparisons of the relative position of items, but the intervals between the categories are not necessarily equal or known. Ordinal data are widely used in surveys, questionnaires, and educational assessments to gauge attitudes, perceptions, and performance levels.

Characteristics of Ordinal Data:

- Categories are distinct and mutually exclusive.
- Categories have a natural order or ranking.
- The intervals between ranks are not necessarily equal.

Examples:

- Education level (high school, bachelor's, master's, PhD)
- Customer satisfaction (very unsatisfied, unsatisfied, neutral, satisfied, very satisfied)
- Engineering Design Maturity: Concept, Prototype, Final Design.
- Performance Ratings: Poor, Fair, Good, Excellent.
- Pain Levels: No Pain, Mild Pain, Moderate Pain, Severe Pain.
- Stage of Disease: Stage I, Stage II, Stage III, Stage IV.
- Quality of Life Scores: Poor, Fair, Good, Excellent.
- Severity of Symptoms: Mild, Moderate, Severe.
- Patient Satisfaction Levels: Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied.

DRAFT

Interval Level

The interval level of measurement involves data that are not only ordered but also have equal intervals between values. Unlike ordinal data, the differences between values are meaningful. However, interval data do not have a true zero point, which means that ratios are not meaningful. Despite lacking a true zero point, interval data provide a higher level of detail compared to nominal and ordinal data, allowing for more sophisticated analytical techniques.

Characteristics of Interval Data:

- Data has a meaningful order.
- The intervals between values are equal and interpretable.
- No true zero point (zero does not mean "nothing").

Examples:

- Temperature Measurements in Celsius or Fahrenheit: 10°C, 20°C, 30°C.
- IQ Scores: Differences between scores are meaningful, but there is no true zero point.
- Calendar Years: The difference between years is consistent, but there is no true zero year.
- Body Temperature in Celsius: 36.5°C, 37.0°C, 37.5°C.
- SAT Scores: There is no true zero score that indicates an absence of knowledge or ability; the score is relative to the testing scale.

Ratio Level

The ratio level of measurement is the highest level of measurement and includes all the properties of the interval level, with the addition of a true zero point. This allows for meaningful comparisons and calculation of ratios. The presence of a true zero point allows for a full range of mathematical and statistical operations, making ratio data the most informative and versatile level of measurement.

Characteristics of Ratio Data:

- Data has a meaningful order.
- The intervals between values are equal and interpretable.
- There is a true zero point (zero means "nothing").
- Ratios are meaningful.

DRAFT

Examples:

- Length of Engineering Components: 5 m, 10 m, 15 m.
- Weight of Materials: 2 kg, 5 kg, 10 kg.
- Cost of Engineering Projects: $1000, $2000, $3000.
- Mechanical Stress Measurements: 10 MPa, 20 MPa, 30 MPa.
- Production Quantity: 100 units, 200 units, 300 units.
- Height of Patients: 150 cm, 160 cm, 170 cm.
- Weight of Patients: 50 kg, 60 kg, 70 kg.
- Number of Hospital Visits: 1 visit, 2 visits, 3 visits.
- Dosage of Medication: 10 mg, 20 mg, 30 mg.
- Length of Hospital Stay: 1 day, 2 days, 3 days.

By grasping these basic statistical concepts, data scientists can better analyze and interpret data, leading to more accurate and meaningful insights. As we delve deeper into data science, these foundational principles will serve as the building blocks for more advanced topics and applications.

1.6 Variables

In statistics and data science, a variable is any characteristic or properties that can take on different values. Variables are essential for research, as they represent the different factors or elements that can change or vary across different individuals, conditions, or time periods. They can take on different values depending on the nature of the data being collected.

> **Variable:** A characteristic or quantity that can vary among individuals, objects, or over time.

Understanding the types of variables is crucial because it determines the kind of statistical analysis that can be performed. Variables are broadly categorized into **qualitative** and **quantitative** variables, each with further subtypes.

Qualitative Variables

Qualitative variables, also known as categorical variables, describe categories or groups. These variables represent characteristics that cannot be measured numerically but can be classified into distinct groups.

DRAFT

- **Nominal Variables**: These are variables with categories that have no natural order or ranking. Examples include gender (male, female), marital status (single, married, divorced), and hair color (blonde, brunette, redhead).

- **Ordinal Variables**: These are variables with categories that have a meaningful order or ranking but the intervals between the categories are not necessarily equal. Examples include education level (high school, bachelor's, master's, doctorate) and customer satisfaction ratings (satisfied, neutral, dissatisfied).

Quantitative Variables

Quantitative variables, also known as numerical variables, represent measurable quantities and can be expressed numerically. They can be further divided into discrete and continuous variables.

- **Discrete Variables**: These variables represent countable values. They take on a finite number of distinct values and are often integers. Examples include the number of students in a class, the number of cars in a parking lot, and the number of books on a shelf.

- **Continuous Variables**: These variables represent measurable quantities that can take on any value within a given range. They are often associated with measurements. Examples include height, weight, temperature, and time.

Understanding the type of variable is essential for choosing the appropriate statistical methods and analyses. Each type of variable provides different insights and requires specific techniques for accurate interpretation and decision-making.

1.7 Scope of Applied Statistics

Applied statistics encompasses a wide range of activities and applications across various fields. Here are some key areas of its scope:

- **Data Collection and Management:** Designing surveys, experiments, and data collection methods.

- **Data Analysis:** Employing statistical techniques to analyze data, including hypothesis testing, regression analysis, and ANOVA.

- **Decision Making:** Providing statistical insights for decision-making in business, healthcare, policy-making, and more.

DRAFT

- **Predictive Modeling:** Developing models to forecast future trends and outcomes based on historical data.
- **Quality Control:** Implementing statistical methods for quality assurance and improvement in manufacturing and services.
- **Market Research:** Analyzing consumer behavior, market trends, and product performance.
- **Epidemiology:** Studying the distribution and determinants of health-related events in populations.
- **Financial Analysis:** Applying statistics to risk assessment, investment strategies, and economic forecasting.
- **Educational Assessment:** Evaluating educational programs, student performance, and teaching effectiveness.

1.8 Statistical Methods in Data Science

Statistical methods play a central role in data science by providing the tools and techniques needed to analyze, interpret, and draw conclusions from data. These methods help data scientists understand underlying patterns, make predictions, and inform decision-making. Below are some key statistical methods commonly used in data science:

- **Descriptive Statistics:** Descriptive statistics involve summarizing and describing the main features of a dataset. This can include measures such as:
  - *Mean:* The sum of all values divided by the number of values.
  - *Median:* The middle value when the data is sorted.
  - *Mode:* The most frequently occurring value.
  - *Standard Deviation:* A measure of how spread out the values are from the mean.
  - *Variance:* The square of the standard deviation, representing the spread of the data.
- **Inferential Statistics:** Inferential statistics allow us to make predictions or inferences about a population based on a sample of data. Key techniques include:
  - *Hypothesis Testing:* Used to test assumptions or claims about a population based on sample data, using tests like t-tests or chi-squared tests.

DRAFT

- *Confidence Intervals:* Estimating the range within which a population parameter is likely to fall, based on sample data.
- *Analysis of Variance (ANOVA):* A method for comparing the means of multiple groups to determine if there are any statistically significant differences between them.

- **Regression Analysis:** Regression techniques are used to model relationships between variables and make predictions. Common methods include:
  - *Linear Regression:* Modeling the relationship between a dependent variable and one or more independent variables.
  - *Logistic Regression:* Used for binary classification problems where the outcome is categorical (e.g., yes/no, success/failure).
- **Probability Theory:** Probability theory helps quantify uncertainty and is fundamental to many data science algorithms. Common tools include:
  - *Bayesian Statistics:* A method of statistical inference that updates probabilities based on new evidence.
  - *Random Variables:* Variables whose values are outcomes of random phenomena, often described by probability distributions (e.g., normal distribution, binomial distribution).
- **Machine Learning Algorithms:** Many machine learning algorithms, especially supervised learning algorithms, are built upon statistical principles. These include:
  - *Classification:* Methods like decision trees, support vector machines (SVM), and neural networks that categorize data into distinct classes.
  - *Clustering:* Techniques like k-means or hierarchical clustering that group data points based on similarity.
  - *Dimensionality Reduction:* Methods such as Principal Component Analysis (PCA) that reduce the number of variables in a dataset while preserving as much information as possible.
- **Time Series Analysis:** Time series methods are used for analyzing data collected over time. Techniques such as:
  - *ARIMA (AutoRegressive Integrated Moving Average):* A model used for forecasting time series data.

DRAFT

    - *Seasonal Decomposition:* Identifying and removing seasonal patterns from time series data to better understand underlying trends.

- **Statistical Sampling:** Sampling techniques are used to select subsets of data from a larger population for analysis. Methods include:
    - *Random Sampling:* Selecting data randomly to ensure unbiased representation of the population.
    - *Stratified Sampling:* Dividing the population into subgroups and sampling from each group to ensure representation from all segments.

1.9 Overview of Data Science Workflow

The data science workflow typically involves several key steps:

1. **Problem Definition**: Understanding the problem and defining objectives.
2. **Data Collection**: Gathering relevant data from various sources.
3. **Data Cleaning and Processing**: Handling missing values, outliers, and ensuring data quality.
4. **Data Exploration**: Summarizing and visualizing data to understand its structure and patterns.
5. **Modeling**: Applying statistical and machine learning techniques to build predictive models.
6. **Evaluation**: Assessing the performance of the models using appropriate metrics.
7. **Deployment**: Implementing the model in a production environment.
8. **Monitoring**: Continuously monitoring and refining the model to ensure its accuracy and relevance.

1.10 Popular Statistical Analysis Tools

There are several tools and software commonly used for statistical analysis in data science, including:

- **Python**: A versatile programming language with powerful libraries such as NumPy, pandas, SciPy, and scikit-learn for statistical analysis and machine learning.

DRAFT

- **R**: A programming language and software environment designed for statistical computing and graphics.
- **SAS**: A software suite used for advanced analytics, multivariate analysis, business intelligence, and data management.
- **STATA**: A powerful software provides comprehensive tools for data manipulation, statistical analysis, and graphical representation.
- **SQL**: A domain-specific language used for managing and manipulating relational databases.
- **Excel**: A spreadsheet software that provides basic statistical functions and data visualization capabilities.

1.11 Why Choose Python for This Book?

Python has become the language of choice for many data scientists and statisticians. Here are several reasons why Python is particularly suited for statistical analysis and data science:

1. Ease of Learning and Use

Python's syntax is straightforward and readable, making it accessible for beginners. Its simplicity allows users to focus on solving problems rather than getting bogged down by the complexities of the language itself.

2. Comprehensive Libraries

Python boasts a rich ecosystem of libraries that are essential for statistical analysis and data science. Some of the most popular libraries include:

- **NumPy**: Provides support for large multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
- **Pandas**: Offers data structures and functions needed to manipulate structured data seamlessly, making it easy to handle and analyze data.
- **SciPy**: Contains modules for optimization, linear algebra, integration, and other advanced mathematical functions.
- **Matplotlib and Seaborn**: Provide powerful tools for data visualization, allowing users to create a wide range of static, animated, and interactive plots.
- **Statsmodels**: Enables statistical modeling, hypothesis testing, and data exploration.

DRAFT

- **Scikit-learn**: A robust library for machine learning that includes simple and efficient tools for data mining and data analysis.

3. Community and Support

Python has a large, active community of users and developers. This means extensive documentation, a wealth of tutorials, and a plethora of forums where users can ask questions and share knowledge. This community support accelerates learning and problem-solving.

4. Integration Capabilities

Python integrates well with other languages and technologies. It can be used alongside other data tools like SQL for database queries, or even integrated with languages such as R or C++ for specialized tasks. This flexibility makes Python a versatile tool in a data scientist's toolkit.

5. Open Source and Free

Python is open source and freely available, which lowers the barrier to entry for individuals and organizations. This democratization of technology allows more people to engage with data science and statistical analysis.

6. Versatility

Python is not only used for data analysis but also for web development, automation, scripting, and even artificial intelligence and machine learning. This versatility means that once you learn Python, you can apply your skills to a wide range of problems and projects.

7. Real-World Applications

Many industry leaders and tech giants like Google, Facebook, and NASA use Python for data analysis and machine learning. This real-world application underscores Python's reliability and effectiveness in handling complex data tasks.

8. Continuous Development

Python and its libraries are continuously being developed and improved by the community, ensuring that users have access to the latest tools and techniques in data science and statistics.

In summary, Python's combination of simplicity, powerful libraries, community support, integration capabilities, and versatility makes it an ideal choice for data scientists and statisticians. Throughout this book, we will leverage

DRAFT

Python to demonstrate various statistical tools and methodologies, ensuring that you can apply what you learn to real-world data challenges effectively.

1.12 Getting Started with Python

Setting Up Python

Instructions for installing Python can be found at `https://www.python.org/downloads/`. We recommend using Jupyter Notebook or an Integrated Development Environment (IDE) like PyCharm.

Basic Python Concepts

If you are new to Python, start with the basics: variables, data types, control structures, and functions. Jupyter Notebooks are particularly useful for interactive data analysis.

1.13 Structure of the Book

Overview of Chapters

This book is structured to gradually build your knowledge in data science, starting from basic concepts to advanced applications. Each chapter includes theoretical explanations, practical examples, and Python code snippets.

Learning Path

To get the most out of this book, follow the chapters sequentially. Practice the examples and exercises provided to reinforce your understanding.

1.14 Concluding Remarks

Applied statistics is a fundamental component of data science that provides the tools necessary for data analysis and interpretation. By understanding basic statistical concepts and methods, data scientists can perform effective analyses and make data-driven decisions. In the following chapters, we will delve deeper into specific statistical methods and explore how they can be applied to real-world data science problems.

By the end of this book, you will have a solid foundation in applied statistics and probability using Python. You will be equipped with the skills to tackle real-world data science problems.

DRAFT

1.15 Chapter Exercises

1. Define data science. How does it differ from traditional data analysis?
2. List and describe the various stages of a data science project lifecycle.
3. Explain the role of a data scientist. What skills are essential for a data scientist to be effective in their role?
4. Discuss the importance of interdisciplinary knowledge in data science. Provide examples of how data science can be applied in different fields.
5. Define statistics. Provide examples of how statistics are applied in the fields of healthcare and engineering.
6. Explain the difference between descriptive and inferential statistics. Give an example of each.
7. Discuss the importance of data visualization in statistical analysis. Mention two Python libraries that are useful for creating data visualizations.
8. Define population and sample. Why is sampling important in statistical analysis?
9. Define parameter and statistic with an example.
10. What is meant by data? Identify the type of data (quantitative or qualitative) for each of the following:
    (i). The colors of cars in a parking lot.
    (ii). The heights of students in a class.
    (iii). The brands of smartphones owned by a group of people.
    (iv). The number of books read by a group of students in a year.
    (v). The types of cuisine served at different restaurants.
11. Define variable. For each of the following variables, identify the level of measurement (nominal, ordinal, interval, or ratio):
    (i). The ranking of movies from a film festival.
    (ii). The temperatures in degrees Celsius recorded over a week.
    (iii). The number of steps taken by an individual in a day.
    (iv). The blood types of patients in a hospital.
    (v). The ages of participants in a survey.
12. Categorize the following data sets as either nominal, ordinal, interval, or ratio:

DRAFT

(i). Survey responses (Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree).

(ii). Birth years of employees in a company.

(iii). Types of pets owned (dog, cat, bird, etc.).

(iv). Test scores out of 100.

(v). Customer satisfaction ratings on a scale of 1 to 10.

DRAFT

Chapter 2

Data Summarization and Visualization

DRAFT

2.1 Introduction

Data exploration is a crucial first step in any data analysis or data science project. It helps us understand the data's structure, distribution, and essential characteristics before proceeding to more advanced analyses. In this chapter, we will learn how to use both tabular and graphical methods to summarize and visualize data effectively.

For **qualitative data**, we will explore techniques such as frequency distribution tables, bar charts, and pie charts. These tools help us capture and display categorical data patterns in a clear, informative manner. With the aid of modern programming languages like Python, we can quickly create these visualizations and gain immediate insights into our data.

For **quantitative data**, we will discuss how to construct frequency distribution tables and visualize data using line diagram, histograms, frequency polygons, ogives, and stem-and-leaf plots. These graphical methods help us identify data concentration, trends, and potential outliers, which are essential for further analysis.

By the end of this chapter, you will have a solid foundation for exploring and summarizing data using both classical statistical techniques and modern programming tools. This will equip you with the confidence to dive deeper into data science projects with a clear understanding of your data's fundamental properties.

2.2 Summarizing Qualitative Data

Qualitative data, also known as **categorical data**, represent characteristics or attributes that can be grouped into categories rather than measured numerically. Examples include gender, colors, or types of movies. **Summarizing Qualitative Data** involves organizing and presenting categorical data in a way that highlights the frequency and distribution of the different categories. A **frequency distribution table** displays the number of observations in each category and provides a clear view of the data structure.

Construction of a Frequency Distribution Table

Constructing a frequency distribution table is a fundamental step in exploring and summarizing data. It involves organizing raw data into a structured format that shows how often each unique category or value appears. Some important steps are listed below:

(i). **List Unique Categories or Classes:** Start by identifying all the distinct categories or classes in your dataset. For example, when analyzing students' favorite type of movies, the categories might include Comedy, Action, Romance, Drama, and Science Fiction.

(ii). **Tally the Frequencies:** For each category, count how many times it appears in the dataset. You can use tally marks or simply count the occurrences directly.

(iii). **Record Frequencies (f_i):** Enter the counts (frequencies) for each category into the table. This provides a clear picture of the data's distribution.

(iv). **Calculate Relative Frequencies (Optional):** If desired, expresses category frequencies as proportions or percentages of the total, allowing comparisons across different datasets. The relative frequency is calculated by using following formula:

$$\text{Relative Frequency} = \frac{\text{Frequency of the class}}{\text{Total number of observations}} = \frac{f_i}{\sum_i f_i} = \frac{f_i}{n}$$

where f_i is the frequency of the ith class and n is the sample size.

To express it as a percentage,

$$\text{Relative Frequency (\%)} = \left(\frac{\text{Frequency of the class}}{\text{Total number of observations}}\right) \times 100\%$$

DRAFT

(v). **Present the Table:** Organize the information in a table with clearly labeled columns for categories, frequencies, and (if applicable) relative frequencies.

This structured approach to organizing data is crucial for both qualitative and quantitative variables. For quantitative data, detailed steps are explained in Section 2.3. This method not only provides a numerical summary but also lays the foundation for creating graphical displays such as bar charts and pie charts.

Problem 2.1. *A survey was conducted among students to determine their favorite type of movie from the following options: Comedy, Action, Romance, Drama, and Science Fiction. Responses were collected from a total of 20 students. The raw data of their preferences is presented in Table 2.1.*

Table 2.1: Students' favorite type of movie (raw data).

Comedy	Science Fiction	Comedy	Comedy
Action	Science Fiction	Action	Romance
Action	Romance	Science Fiction	Romance
Romance	Romance	Action	Drama
Comedy	Romance	Action	Science Fiction

(a). Construct a frequency distribution table for the type of movie preferences.

(b). Comment on the students' preferences based on the frequency distribution.

Solution:

(a). The favorite type of movies of the 20 students are organized into the following frequency distribution table 2.2.

DRAFT

Table 2.2: Frequency distribution of movie types among students preferences

Category	**Tally** Marks	**Frequency** (f_i)	**Relative Frequency (%)**
Comedy	$\|\|\|\|$	4	20%
Action	$\cancel{\|\|\|\|}$	5	25%
Romance	$\cancel{\|\|\|\|}\,\|$	6	30%
Drama	$\|$	1	5%
Science Fiction	$\|\|\|\|$	4	20%
Total		$\sum_i f_i = 20$	100%

(b). Comment on the students' preferences:

- *Romance* is the most preferred type of movie (30% of students).
- *Action* is the second most popular (25%).
- *Comedy* and *Science Fiction* are equally preferred (20% each).
- *Drama* is the least preferred (5%).
- Overall preferences are fairly spread across type of movies, with a modest concentration toward Romance and Action.

□

2.2.1 Bar Chart

A **bar chart** (or bar graph) is a type of data visualization used to represent and compare different categories of data through rectangular bars. The length or height of each bar is proportional to the frequency or relative frequency of the corresponding category. Bar charts provide an effective way to compare different categories visually.

Bar Chart: A chart that uses rectangular bars to represent the frequency or proportion of categories in a dataset, with the length of each bar corresponding to its value.

Construction of a Bar Chart

The steps to construct a bar chart are as follows:

(i). **Organize the data into a frequency distribution table.** List all categories and their corresponding frequencies or counts.

DRAFT

(ii). **Choose an appropriate scale.** Determine the scale for the vertical axis based on the range of frequencies. Ensure the scale is clear and consistent.

(iii). **Draw the axes.** On graph paper or using software, draw two perpendicular axes. Label the horizontal axis with the category names and the vertical axis with the frequency scale.

(iv). **Draw the bars.** For each category, draw a bar whose height (or length, if horizontal) corresponds to its frequency. Bars should be of equal width and spaced evenly.

(v). **Label the chart.** Add a title, and label both axes clearly. Optionally, include frequency values or percentages on top of each bar for clarity.

Bar charts are widely used because they clearly depict differences among categories and can easily accommodate large numbers of categories without losing readability. It is a powerful tool for visualizing categorical data. It allows for easy comparison between different categories. This makes it straightforward to identify patterns, trends, and outliers within the data.

Problem 2.2. *Refer to Table 2.1 for the dataset needed to create a bar chart. Using the dataset provided in Table 2.2, construct a bar chart to display the distribution of students' favorite type of movies. Based on the bar chart, describe and interpret the main patterns in the data.*

Solution: To draw the **bar chart**, the frequency distribution of the student's preferences is given in Table 2.2. The bar chart is presented in Figure 2.1.

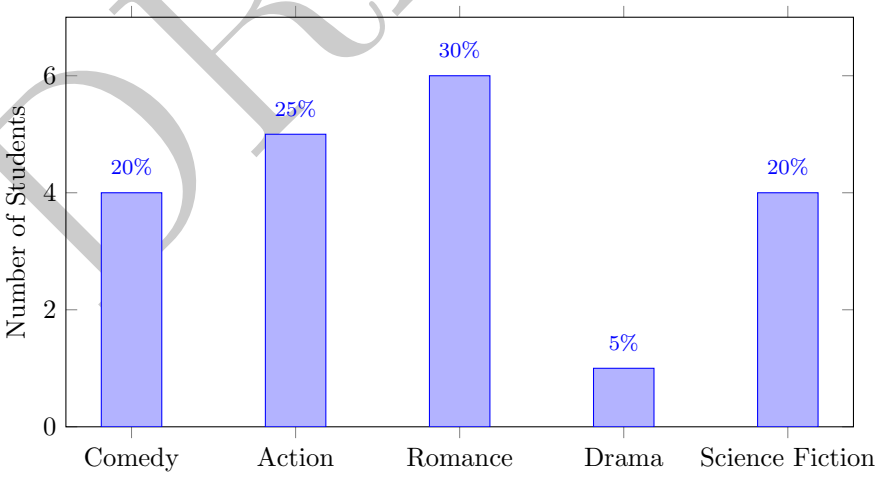


Figure 2.1: Bar Chart of Movie Preferences

DRAFT

Comments:
The chart shows that *Romance* is the most popular genre, preferred by 6 students (30%), followed closely by *Action*, chosen by 5 students (25%). Both *Comedy* and *Science Fiction* share equal popularity, each selected by 4 students (20%). In contrast, *Drama* is the least favored, with only 1 student (5%) indicating it as their choice. Overall, more than half of the students (55%) prefer either Romance or Action, while the remaining 45% are distributed among Comedy, Science Fiction, and Drama.

□

2.2.2 Python Code: Bar Chart

To create a bar chart for the data in Example 2.1 using Python, you can use the following code to generate the chart.

```
#Install the necessary libraries
import matplotlib.pyplot as plt
from collections import Counter

# Data from the table
movies = [
"Comedy", "Science Fiction", "Comedy", "Comedy",
"Action", "Science Fiction", "Action", "Romance",
"Action", "Romance", "Science Fiction", "Romance",
"Romance", "Romance", "Action", "Drama",
"Comedy", "Romance", "Action", "Science Fiction"
]

# Count the frequency of each movie genre
movie_counts = Counter(movies)

# Extract the genres and their corresponding counts
genres = list(movie_counts.keys())
counts = list(movie_counts.values())

# Plotting the bar chart
plt.figure(figsize=(10, 6))
plt.bar(genres, counts, color='skyblue')
plt.xlabel('Movie Genre')
plt.ylabel('Frequency')
plt.title('Favorite Movie Genres')
plt.show()
```

2.2.3 Pie Chart

A **pie chart** is a circular statistical graphic divided into slices to illustrate numerical proportions. Each slice of the pie represents a category's contribution to the total, with the entire pie representing 100% of the data.

Pie Chart: A circular chart that represents the proportion of each category in a dataset as slices, allowing easy comparison of relative frequencies.

Pie charts are a straightforward way to show how different parts contribute to a whole, making them a popular choice for visualizing proportions in data analysis.

How to Construct a Pie Chart?

The process to construct a pie chart involves the following steps:

(i). **Prepare a frequency distribution table.** Begin by listing all the categories of interest along with their corresponding frequencies or counts.

(ii). **Calculate the relative frequencies.** Convert the frequencies into proportions or percentages by dividing each frequency by the total number of observations.

(iii). **Determine the angle for each slice.** Multiply each relative frequency by 360 degrees to obtain the angle size for each sector of the pie chart:

$$\begin{aligned}\text{Angle for each category} &= \text{Relative frequency} \times 360^\circ \\ &= \frac{f_i}{\sum_i f_i} \times 360^\circ\end{aligned}$$

(iv). **Draw the pie chart.** Using a compass or drawing tool, sketch a circle representing the entire dataset.

(v). **Mark and draw slices.** Starting at the top (or any chosen point), use a **protractor** to measure each angle and draw the slices sequentially, moving clockwise or counterclockwise around the circle.

(vi). **Label the slices.** Clearly label each slice with the corresponding category name. Including the percentage or frequency values enhances the chart's interpretability.

Pie charts are especially effective when the number of categories is limited and the goal is to emphasize the proportional relationships within the data. They offer a quick visual impression of how the parts relate to the whole.

Problem 2.3. *Refer to Table 2.1 for the dataset needed to create a pie chart. Analyze the pie chart and interpret the key findings from the data visualized.*

DRAFT

Solution: To draw the **pie chart**, the frequency distribution of the student's preferences is given in Table 2.2. The pie chart is presented in Figure 2.2.

Table 2.3: Distribution of Favorite Movie Types Among Students

Category	Frequency (f_i)	Percentage	Angle = $\frac{f_i}{\sum_i f_i} \times 360°$
Comedy	4	20%	72°
Action	5	25%	90°
Romance	6	30%	108°
Drama	1	5%	18°
Science Fiction	4	20%	72°
Total	$\sum_i f_i = 20$		

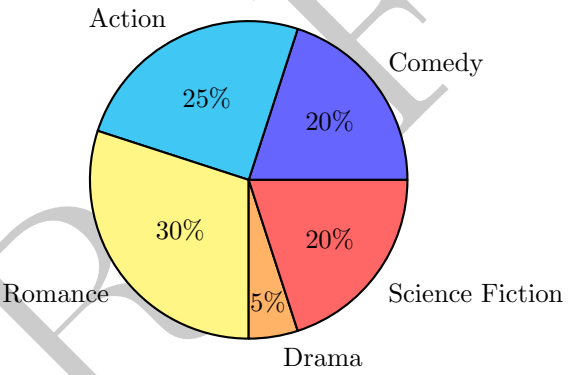


Figure 2.2: Pie Chart of Movie Preferences

The frequency distribution table and pie chart reveals the preferences for type of movies among 20 respondents. Romance emerges as the most preferred genre, accounting for 30% of the choices, as indicated by the largest pie chart segment. Action is also quite popular, chosen by 25% of respondents. Both Comedy and Science Fiction have equal popularity, each being preferred by 20% of the respondents. Drama is the least favored genre, with only 5% of the respondents selecting it. These results suggest that respondents have a diverse range of genre preferences, with Romance standing out as the most popular choice.

DRAFT

2.2.4 Python Code: Pie Chart

To create a bar chart for the data in Example 2.1 using Python, you can use the following code to generate the chart.

```
pip install pandas matplotlib # Install Required Libraries (if not already installed)
import matplotlib.pyplot as plt
from collections import Counter

# Data from the table
data = [
'Comedy', 'Science Fiction', 'Comedy', 'Comedy',
'Action', 'Science Fiction', 'Action', 'Romance',
'Action', 'Romance', 'Science Fiction', 'Romance',
'Romance', 'Romance', 'Action', 'Drama',
'Comedy', 'Romance', 'Action', 'Science Fiction'
]

# Count occurrences of each genre
genre_counts = Counter(data)

# Prepare data for the pie chart
genres = list(genre_counts.keys())
counts = list(genre_counts.values())

# Create the pie chart
plt.figure(figsize=(10, 7))
plt.pie(counts, labels=genres, autopct='%1.1f%%', startangle=140)
plt.title('Favorite Movies Pie Chart')
plt.show()
```

Problem 2.4. *Imagine you conducted a survey to find out the types of physical activities performed by patients in a rehabilitation program. You collected data from 100 patients, and the results are as follows:*

- *Walking: 25 patients*
- *Cycling: 20 patients*
- *Swimming: 15 patients*
- *Yoga: 15 patients*
- *Strength Training: 10 patients*
- *Pilates: 8 patients*
- *Dancing: 7 patients*

Answer the following questions:

(a). *What is the name of the variable under study? Is this a qualitative variable? If the answer is no, why?*

(b). *Make a pie chart to represent the distribution of physical activities among the patients. Write a summary based on this. Be sure to label the chart accurately and show the percentage of patients for each activity.*

Solution:

(a). The variable under study is the type of physical activity performed by patients in the rehabilitation program. Yes, this is a qualitative variable because it describes categories or types of activities rather than numerical values.

(b). Pie Chart: To make a pie chart, we need to calculate the percentage of patients who performed each activity out of 100 total patients. The percentages are presented in Table 2.4.

Table 2.4: Frequency distribution of physical activities among rehabilitation patients

Activity	**Number of Patients**	**Percentage (%)**
Walking	25	25.0%
Cycling	20	20.0%
Swimming	15	15.0%
Yoga	15	15.0%
Strength Training	10	10.0%
Pilates	8	8.0%
Dancing	7	7.0%
Total	100	100.0%

DRAFT

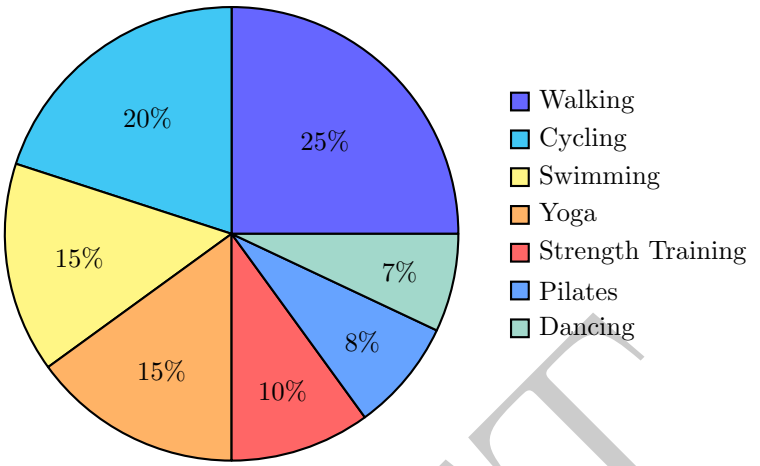


Figure 2.3: Pie chart of physical activities among rehabilitation patients

The pie chart shows the distribution of physical activities among the patients in the rehabilitation program. Walking is the most common activity, with 25% of the patients participating in it. This is followed by cycling (20%), swimming (15%), and yoga (15%). Strength training is performed by 10% of the patients, while pilates and dancing are the least common activities, with 8% and 7% participation, respectively.

2.3 Summarizing Quantitative Data

Summarizing quantitative data involves organizing and presenting numerical data in a way that reveals patterns, trends, and important characteristics of the dataset. Some common methods are explained in the following.

2.3.1 Line Diagram

A **line diagram** (or line graph) is a simple yet powerful graphical tool used to represent discrete quantitative data. It displays individual data points as dots (or markers), which are then connected by straight line segments. This type of diagram helps in visualizing the changes or patterns in data across a sequence, typically over time or ordered categories.

> A **line diagram** is a graph that uses straight lines to connect individual data points, showing trends over time or sequential categories.

The structure of a line diagram can be understood through the following key components:

DRAFT

- **Horizontal axis (x-axis)**: Represents the categories or time intervals, such as months, years, or test numbers.
- **Vertical axis (y-axis)**: Displays the frequencies or values corresponding to each category on the x-axis.
- **Data points and lines**: Individual data values are plotted as points on the graph. These points are then connected using straight lines to form the line diagram.

Advantages of Line Diagram

- Easy to interpret trends and changes over time.
- Clear comparison between multiple data sets (using multiple lines).
- Simple to construct for time series or sequential data.
- Effective for continuous data visualization.

Limitations of Line Diagram

- Not appropriate for continuous grouped data.
- Hard to read with too many lines or data points.
- Interpolation may be misleading, lines may suggest continuity that does not exist.

Problem 2.5. *The table below shows the monthly sales (in thousands) for a company over five months:*

Month	***Sales (in 1000s)***
January	*120*
February	*135*
March	*150*
April	*130*
May	*160*

*Draw a **line diagram** to represent the monthly sales data.*

Solution: To draw a line diagram, plot the months on the horizontal axis and the sales (in thousands) on the vertical axis. Then, mark the sales values for each month and connect the points with straight lines.

DRAFT

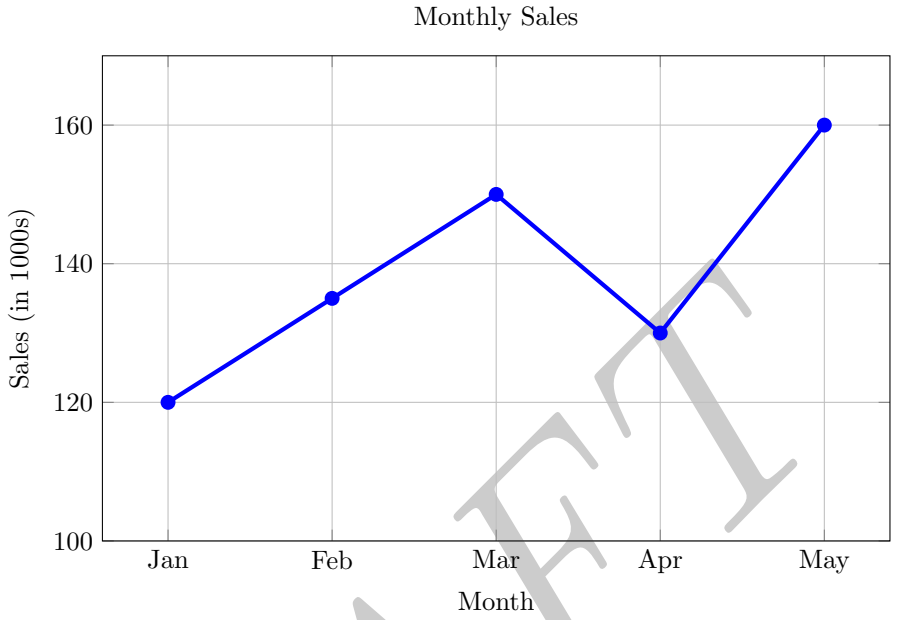


Figure 2.4: Line Diagram of Monthly Sales

The line diagram given in Figure 2.4 shows monthly sales from January to May. Sales increased steadily from January (120000) to March (150000), indicating strong early growth. In April, there was a noticeable dip to 130000, followed by a sharp rise in May to 160000, the highest in the period. Despite the temporary drop, the overall trend is upward. This diagram effectively highlights the fluctuation and recovery in sales over time, making it a useful tool for identifying patterns and supporting data-driven decisions. □

Python Code Example (Using `matplotlib`)

```
import matplotlib.pyplot as plt

months = ['Jan', 'Feb', 'Mar', 'Apr', 'May']
sales = [120, 135, 150, 130, 160]

plt.plot(months, sales, marker='o', linestyle='-', color='
  blue')
plt.title('Monthly Sales')
plt.xlabel('Month')
plt.ylabel('Sales (in 1000s)')
plt.grid(True)
plt.show()
```

DRAFT

Problem 2.6. *The following table shows the runs scored by a cricket team in each over during the first 10 overs of a match:*

Over	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Runs Scored	*6*	*4*	*8*	*3*	*5*	*7*	*2*	*9*	*4*	*6*

(a). Calculate the cumulative runs after each over.

(a). Draw a line diagram showing the cumulative runs scored over the 10 overs.

Solution: *To understand how a cricket team's score progresses over the course of a match, we can use a* ***line diagram*** *that shows the cumulative runs scored after each over. Let's solve the problem step by step.*

(a). *The first task is to find the* ***cumulative runs*** *after each over. Cumulative runs mean the total runs scored up to that point in the match. Given the runs scored in each over:*

Over	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Runs Scored	*6*	*4*	*8*	*3*	*5*	*7*	*2*	*9*	*4*	*6*

We add these runs progressively:

Over	1	2	3	4	5	6	7	8	9	10
Cumulative Runs	6	10	18	21	26	33	35	44	48	54

(b). *Now, we plot a* ***line diagram*** *to visualize the scoring pattern.*

- *The* ***x-axis*** *represents the overs (from 1 to 10).*
- *The* ***y-axis*** *represents the cumulative runs.*
- *Each point on the graph shows the total runs scored after a particular over.*
- *By joining these points with straight lines, the diagram reveals the trend of the team's scoring as the innings progresses.*

DRAFT

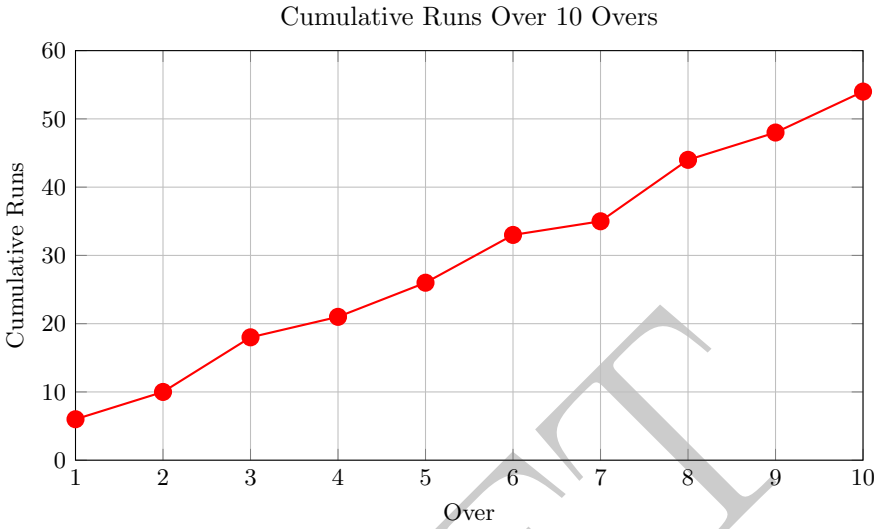


Figure 2.5: Line Diagram of Cumulative Runs Over 10 Overs

Interpretation

- *The slope of the line between points indicates the scoring speed.*
- *Steeper upward segments (like between overs 2 and 3, or 7 and 8) indicate rapid scoring.*
- *Flatter segments (like between overs 6 and 7) indicate slower scoring or a pause.*
- *The overall upward trend confirms the team is steadily accumulating runs.*

2.3.2 Frequency Distribution Table

The general steps to construct a frequency distribution table for quantitative data are the following:

(i). **Identify the minimum and maximum data values.** Determine the smallest and largest observations in the dataset.

(ii). **Determine the number of classes** (k). This can be done in different ways:

- **Sturges' Rule:** This logarithmic formula is widely used to determine the number of classes:

$$k = 1 + \log_2(n),$$

DRAFT

where n is the total number of observations. Using the change-of-base formula,

$$\log_2(n) = \frac{\log_{10}(n)}{\log_{10}(2)} \approx 3.322 \log_{10}(n),$$

we get,

$$k = 1 + 3.322 \log_{10}(n).$$

- Alternatively, choose the smallest integer k that satisfies:

$$2^k \geq n.$$

(iii). **Determine the class interval width** (h). This can be calculated by:

$$h \geq \frac{\text{Maximum value} - \text{Minimum value}}{\text{Number of classes } (k)}.$$

Round h to a convenient and practical value to ensure meaningful and readable class intervals.

(iv). **Define the lower limit of the first class.** Choose a lower limit that covers the minimum value in the data. Define subsequent class intervals to be mutually exclusive and collectively exhaustive, ensuring no overlap and no missing data.

(v). **Tally the data.** Review the dataset and tally how many observations fall within each class interval.

(vi). **Count the frequencies.** Convert tally marks into numerical frequencies for each class interval.

(vii). **Calculate relative frequencies (optional).** To express frequencies as proportions or percentages, divide each class frequency by the total number of observations.

(viii). **Present the table.** Organize the class intervals, frequencies, and (if applicable) relative frequencies into a clear and well-formatted table.

Remark 2.3.1. *Note that the "**Tally**" marks are a helpful intermediate step in manual calculations but not mandatory for constructing or interpreting a frequency distribution. Once frequencies are counted, tallies are often removed or not shown in final tables.*

Let us take the following example to understand how to construct a frequency distribution.

DRAFT

Problem 2.7. *Suppose we have the weekly expenditure (in dollars) of 30 students, as shown below:*

423	369	387	411	393	394
371	377	389	409	392	407
431	401	363	391	405	382
400	381	399	415	428	422
395	371	410	419	386	390

Construct a frequency distribution table using an appropriate number of classes and class intervals.

Solution: We shall construct a frequency distribution table for the given weekly expenditure data, following the systematic steps for quantitative data.

(i). **Identify the minimum and maximum data values.** The minimum value in the dataset is 363, and the maximum value is 431.

(ii). **Determine the number of classes (k).** Using Sturges' rule:

$$k = 1 + 3.322 \log_{10}(30) \approx 1 + 3.322 \times 1.4771 = 5.907 \approx 6.$$

Thus, we select $k = 6$ classes.

(iii). **Determine the class width (h).** The class width is obtained as:

$$h \geq \frac{\text{Maximum value} - \text{Minimum value}}{k} = \frac{431 - 363}{6} = \frac{68}{6} \approx 11.33.$$

We round h up to a convenient number, choosing $h = 12$.

(iv). **Decide the lower limit of the first class.** We take 360 as the lower limit of the first class, slightly below the minimum value, to ensure that the first class includes the smallest observation.

(v). **Tally the data.** Each observation is placed into its corresponding class interval using tally marks.

(vi). **Count the frequencies.** The total number of observations in each class interval is recorded as the class frequency. The intervals and their corresponding frequencies are presented in Table 2.5.

(vii). **Compute relative frequencies (optional).** The relative frequency of each class is computed by dividing the class frequency by the total number of observations ($n = 30$).

(viii). **Present the frequency distribution table.** The complete frequency distribution table is shown below.

DRAFT

Table 2.5: Frequency distribution of weekly expenditures of 30 students.

Class Interval	**Tally**	**Frequency (f_i)**	**Relative Frequency (rf_i)**
360–372	𝍸	4	0.1333
372–384	𝍸	3	0.1000
384–396	𝍸 𝍸	9	0.3000
396–408	𝍸	5	0.1667
408–420	𝍸	5	0.1667
420–432	𝍸	4	0.1333
Total		**30**	**1.0000**

This frequency distribution organizes the weekly expenditure data into six mutually exclusive class intervals. It provides a clear summary of how the expenditures are distributed among the students and forms the basis for graphical representations such as histograms and frequency polygons.

2.3.3 Open-Ended Classes in Frequency Distributions

After learning how to construct a frequency distribution, it's important to understand a special type of class: the **open-ended class**. In some cases, especially with very small or very large data values, it is convenient to group extreme values into classes without clearly defined lower or upper limits.

An **open-ended class** is a class interval that does not have either a lower or upper boundary.

- **Lower open-ended class:** No lower boundary (e.g., "Under 30").
- **Upper open-ended class:** No upper boundary (e.g., "70 and over").

Why Use Open-Ended Classes?

Open-ended classes are used in frequency distributions for the following reasons:

(i). **Accommodate Outliers:** They effectively group extremely high or low values (outliers) without creating numerous empty or very wide class intervals.

(ii). **Enhance Readability:** They simplify the table by concisely representing sparse data at the tails of the distribution.

DRAFT

(iii). **Protect Privacy:** They can obscure exact sensitive values in extreme ranges (e.g., very high incomes).

(iv). **Practicality:** They are useful when precise extreme values are unknown, impractical to measure, or not critical for analysis.

Problem 2.8. *The following dataset represents the ages (in years) of 30 participants in a community fitness program:*

$$15, 22, 18, 25, 30, 35, 28, 40, 42, 45, 48, 50, 55, 58, 60,$$
$$62, 65, 45, 70, 45, 60, 45, 55, 95, 40, 45, 60, 65, 50, 125$$

(a). Determine suitable class intervals and construct a frequency distribution table for the given data.

(b). Ensure that the first and last classes are ***open-ended*** *to accommodate the extreme ages.*

(c). Present your frequency distribution table with clearly labeled columns.

(d). Calculate the relative frequencies for each class, expressed as percentages.

Solution

We are given the ages of 30 participants:

$$15, 22, 18, 25, 30, 35, 28, 40, 42, 45, 48, 50, 55, 58, 60,$$
$$62, 65, 45, 70, 45, 60, 45, 55, 95, 40, 45, 60, 65, 50, 125$$

Following the steps for constructing a frequency distribution:

(a) Determine Suitable Class Intervals: The smallest age is 15 and the largest is 125. Using Sturges' Rule:

$$k = 1 + 3.322 \log_{10}(30) \approx 1 + 3.322 \times 1.477 = 5.9 \approx 6 \text{ classes.}$$

Determine the class interval width (h). Calculate the class interval (h):

$$h \geq \frac{125 - 15}{6} \approx 18.33.$$

We round h to a convenient value, 20, for easier class intervals.

(b) Open-Ended Classes: To handle the extreme ages in the dataset, we define open-ended classes for the first and last intervals:

- First class: less than 30
- Last class: 110 and above

DRAFT

(c) Construct the Frequency Distribution Table: Using a class width of 20 and the given boundaries, the class intervals are: Less than 30, 30–50, 50– <70, 70– <90, 90– <110, 110 and above

The frequency distribution table is constructed as follows:

Class Interval	**Frequency (f_i)**	**Relative Frequency**
Less than 30	5	0.167
$30- <50$	11	0.367
$50- <70$	11	0.367
$70- <90$	1	0.033
$90- <110$	1	0.033
110 and above	1	0.033
Total	**30**	**1.00**

(d) Calculate the Relative Frequencies: The relative frequency for each class interval is obtained by using following formula:

$$\text{Relative Frequency} = \frac{\text{Frequency of Class}}{\text{Total Number of Observations}} \times 100$$

Applying this formula to each class interval, the relative frequencies are as follows:

Class Interval	**Relative Frequency (%)**
<30	16.7%
$30- <50$	36.7%
$50- <70$	36.7%
$70- <90$	3.3%
$90- <110$	3.3%
110+	3.3%
Total	**100%**

DRAFT

2.3.4 Python Code: Frequency Distribution

To create a frequency distribution table for the data in Problem 2.7 using Python, you can use the following code to generate the Frequency Distribution table.

```
import pandas as pd
import matplotlib.pyplot as plt

```

```
# Data: weekly expenditure of 30 students
data = [
423, 369, 387, 411, 393, 394,
371, 377, 389, 409, 392, 407,
431, 401, 363, 391, 405, 382,
400, 381, 399, 415, 428, 422,
395, 371, 410, 419, 386, 390
]

# Convert the data into a pandas DataFrame
df = pd.DataFrame(data, columns=['Expenditure'])

# Define the class intervals (bins)
bins = range(360, 440, 12)  # Create bins from 360 to 440 with an interval of 12
labels = [f'{bins[i]}-{bins[i+1]-1}' for i in range(len(bins)-1)]  # Labels for each bin

# Bin the data and calculate frequency distribution
df['Bins'] = pd.cut(df['Expenditure'], bins=bins, labels=labels, right=False)
frequency_distribution = df['Bins'].value_counts().sort_index()

# Print frequency distribution
print("Frequency Distribution:")
print(frequency_distribution)

# Plot frequency distribution as a bar chart
plt.figure(figsize=(12, 6))
frequency_distribution.plot(kind='bar', color='skyblue')
plt.xlabel('Expenditure Range')
plt.ylabel('Frequency')
plt.title('Frequency Distribution of Weekly Expenditure with Fixed Class Intervals')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()


```

2.3.5 Frequency Polygon

A **frequency polygon** is a graphical representation of a frequency distribution. It is constructed by plotting the midpoints of each class interval on the horizontal axis and the corresponding frequencies on the vertical axis. These points are then connected by straight lines to form a polygon.

To draw a frequency polygon, follow these steps.

(i). Calculate the midpoint (x_i) of each class interval using the formula:

$$\text{Midpoint } (x_i) = \frac{\text{Lower class limit} + \text{Upper class limit}}{2}$$

(ii). Plot the midpoints (x_i) on the horizontal axis (x-axis) against their respective frequencies (f_i) on the vertical axis (y-axis).

(iii). Connect the points with straight lines.

(iv). Optionally, extend the polygon to the horizontal axis at both ends to close the figure.

The frequency polygon is useful for comparing different distributions and for visualizing the shape of the data, such as skewness and modality.

A **frequency polygon** is a line graph that displays the frequencies of different class intervals. It is created by plotting the frequencies of each class interval at the midpoints of those intervals and connecting these points with straight lines.

Problem 2.9. *Consider Problem 2.7. Draw frequency polygon of weekly expenditure of 30 students.*

Solution: Consider Problem 2.7. The midpoints for each class interval are calculated and presented in Table 2.6. Then, the midpoints are plotted on the x-axis and the corresponding frequencies on the y-axis. The resulting frequency polygon is depicted in Figure 2.6.

Table 2.6: Distribution of Weekly Expenditure of 30 Students

Class Interval	**Tally**	**Frequency** (f_i)	**Relative Frequency**	**Midpoint** (x_i)
360 - 372	𝍸	4	0.1333	366
372 - 384	𝍸	3	0.1000	378
384 - 396	𝍸 𝍸	9	0.3000	390
396 - 408	𝍸	5	0.1667	402
408 - 420	𝍸	5	0.1667	414
420 - 432	𝍸	4	0.1333	426
Total		30	1.0000	

DRAFT

In the frequency polygon, the peak at the midpoint of 390 indicates that most students' expenditures cluster around this value.The shape of the polygon, with a rise to the peak and a gradual decline, shows that while expenditures are concentrated in the 384–396 range, there is a moderate spread across other ranges. This visualization helps to quickly understand the central tendency and variability of expenditures among the students.

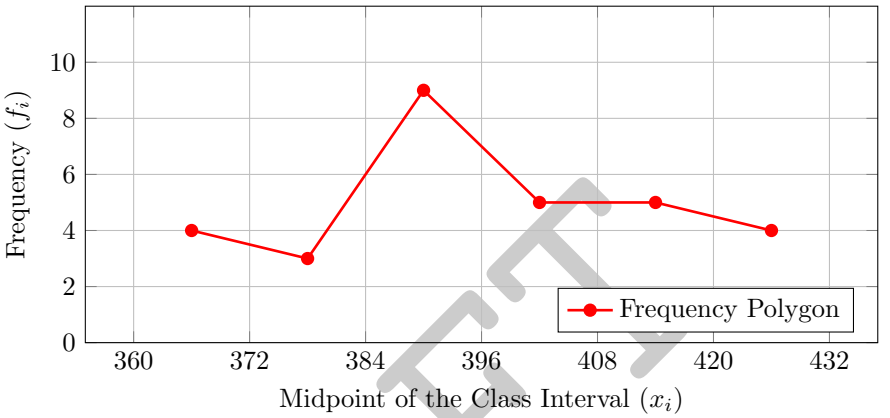


Figure 2.6: Frequency Polygon of Weekly Expenditure of 30 Students

2.3.6 Python Code: Frequency Polygon

To create a frequency polygon for the data in Example 2.7 using Python, you can use the following code to generate the Frequency Polygon.

```
import pandas as pd
import matplotlib.pyplot as plt

# Data: weekly expenditure of 30 students
data = [
423, 369, 387, 411, 393, 394,
371, 377, 389, 409, 392, 407,
431, 401, 363, 391, 405, 382,
400, 381, 399, 415, 428, 422,
395, 371, 410, 419, 386, 390
]

# Convert the data into a pandas DataFrame
df = pd.DataFrame(data, columns=['Expenditure'])

# Define the class intervals (bins)
bins = range(360, 440, 12)  # Create bins from 360 to 440
    with an interval of 12
```

DRAFT

```
labels = [f'{bins[i]}-{bins[i+1]-1}' for i in range(len(bins)-1)]  # Labels for each bin

# Bin the data and calculate frequency distribution
df['Bins'] = pd.cut(df['Expenditure'], bins=bins, labels=labels, right=False)
frequency_distribution = df['Bins'].value_counts().sort_index()

# Calculate bin midpoints
bin_midpoints = [(bins[i] + bins[i+1]) / 2 for i in range(len(bins)-1)]

# Plot frequency polygon
plt.figure(figsize=(12, 6))
plt.plot(bin_midpoints, frequency_distribution.sort_index(), marker='o', linestyle='-', color='b')
plt.xlabel('Expenditure Range')
plt.ylabel('Frequency')
plt.title('Frequency Polygon of Weekly Expenditure')
plt.xticks(bin_midpoints, labels=labels, rotation=45)
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()


```

2.3.7 Ogive Curve

An **ogive curve**, also known as a cumulative frequency polygon, is a graphical representation used in statistics to display the cumulative frequency distribution of a dataset. It shows how many data values fall below or at a certain point, helping to visualize cumulative totals and identify key percentiles.

To construct an ogive curve, follow these steps:

(i). Calculate the cumulative frequency for each class interval which is calculated by adding the frequency of the current class interval to the cumulative frequency of the previous class interval.

(ii). Plot the upper class boundary of each interval on the horizontal axis and the corresponding cumulative frequency on the vertical axis.

(iii). Connect the plotted points with a smooth curve or straight lines.

The ogive is useful for determining median, quartiles, percentiles, and understanding the overall distribution of the data.

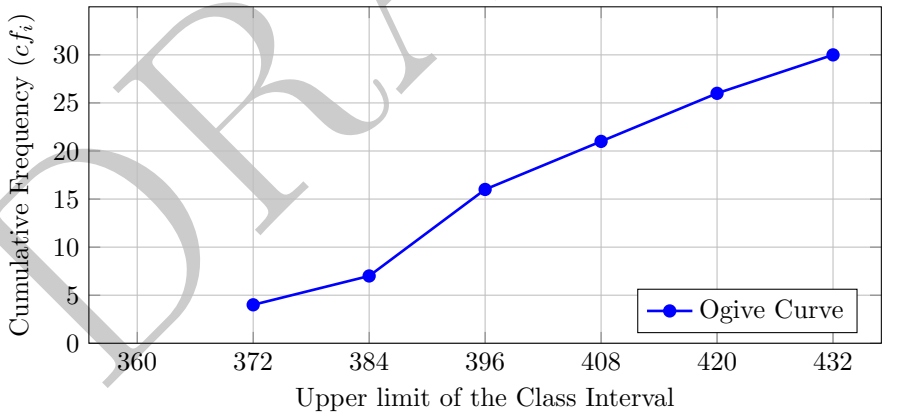
Problem 2.10. *Consider Problem 2.7. Draw Ogive Curve of weekly expenditure of 30 students.*

Solution: Consider Problem 2.7. Using the data from Table 2.6, we calculate cumulative frequency for each class. The frequency distribution table with cumulative frequency is presented in Table 2.7. Use the upper boundaries of each class to plot against cumulative frequencies.

Table 2.7: Distribution of Weekly Expenditure of 30 Students

Class Interval	**Tally**	**Frequency** (f_i)	**Relative Frequency**	**Midpoint** (x_i)	**Cumulative Frequency** (cf_i)
360 - 372	𝍸	4	0.1333	366	4
372 - 384	𝍸	3	0.1000	378	7
384 - 396	𝍸 𝍸	9	0.3000	390	16
396 - 408	𝍸	5	0.1667	402	21
408 - 420	𝍸	5	0.1667	414	26
420 - 432	𝍸	4	0.1333	426	30
Total		30	1.0000		

Figure 2.7: Ogive Curve of Weekly Expenditure of 30 Students



In the example of weekly expenditures for 30 students, the ogive curve given in Figure 2.7, illustrates that as expenditure increases, the cumulative number of students also rises. The curve starts at the cumulative frequency of 4 for the interval 360-372 and gradually increases to 30 for the interval 420-432, reflecting that 30 students' expenditures are up to 432. The steepness of the curve indicates intervals with higher frequencies, while flatter sections show

DRAFT

lower frequencies. The ogive provides insights into data distribution, helping to visualize the proportion of students spending up to various amounts.

2.3.8 Python Code: Ogive

To create a Ogive for the data in Example 2.7 using Python, you can use the following code to generate the Ogive.

```
import pandas as pd
import matplotlib.pyplot as plt

# Data: weekly expenditure of 30 students
data = [
423, 369, 387, 411, 393, 394,
371, 377, 389, 409, 392, 407,
431, 401, 363, 391, 405, 382,
400, 381, 399, 415, 428, 422,
395, 371, 410, 419, 386, 390
]

# Convert the data into a pandas DataFrame
df = pd.DataFrame(data, columns=['Expenditure'])

# Define the class intervals (bins) with an interval of 12
bin_start = 360
bin_end = 440
bin_interval = 12
bins = list(range(bin_start, bin_end + bin_interval,
    bin_interval))

# Generate labels for the bins
labels = [f'{bins[i]}-{bins[i+1]-1}' for i in range(len(bins
    )-1)]

# Bin the data and calculate frequency distribution
df['Bins'] = pd.cut(df['Expenditure'], bins=bins, labels=
    labels, right=False)
frequency_distribution = df['Bins'].value_counts().
    sort_index()

# Calculate cumulative frequency
cumulative_frequency = frequency_distribution.cumsum()

# Calculate bin edges for plotting
bin_edges = bins  # Use only the bin edges

# Calculate cumulative frequencies including the starting
    point
```

```
cumulative_frequency_with_start = [0] + list(cumulative_frequency)

# Ensure the length of bin_edges matches cumulative_frequency_with_start
if len(bin_edges) != len(cumulative_frequency_with_start):
# Extend bin_edges to match the length of cumulative_frequency_with_start
bin_edges = bins + [bins[-1] + bin_interval]

# Plot the ogive
plt.figure(figsize=(12, 6))
plt.plot(bin_edges, cumulative_frequency_with_start, marker='o', linestyle='-', color='b')
plt.xlabel('Expenditure Range')
plt.ylabel('Cumulative Frequency')
plt.title('Ogive of Weekly Expenditure with Interval 12')
plt.xticks(bin_edges, labels=[f'{edges}-{edges+12-1}' for edges in bin_edges[:-1]] + [''])
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

2.3.9 Histogram

A **histogram** is a powerful and intuitive graphical representation of the distribution of numerical data. It is a cornerstone visualization in a data scientist's toolkit, displaying the distribution of a numerical variable by dividing the data into a series of contiguous, non-overlapping intervals (called bins) and then counting how many data points fall into each bin. The height of each bar represents the frequency (or sometimes the proportion) of observations within that bin. This provides a visual summary of the data's distribution, helping to identify patterns such as trends, peaks, and the spread of the data.

> A **histogram** is a graphical representation of the distribution of numerical data, showing the frequency or relative frequency of observations within contiguous, non-overlapping intervals (bins).

The term "histogram" was coined by the British mathematician and statistician **Karl Pearson**. While Pearson formalized the term and popularized the visualization, similar graphical representations of data distribution existed before his time. Nevertheless, he is widely credited for giving the visualization its specific name and for his significant contributions to the field of mathematical statistics.

DRAFT

Steps for Constructing a Histogram

Having established the frequency distribution of the quantitative data, we can now construct a histogram to visually represent it. Each class interval (or bin) from the frequency table is depicted as a contiguous bar whose width matches the interval's range and whose height reflects the frequency of observations within that interval. This adjacency of bars underscores the continuous nature of the data.

To draw a histogram, follow these steps:

(i). **Construct Frequency Distribution:** Construct a frequency distribution table using an appropriate class width (bin) as the first step of the histogram drawing process. The detailed steps for creating a frequency distribution table are explained in Section 2.3.2.

(ii). **Choose Appropriate Scales for the Axes:** Draw the horizontal axis to represent the class intervals and the vertical axis to show the frequencies (or relative frequencies). Ensure the scale is even and suitable for the range of frequencies (or relative frequencies).

(iii). **Plot the Bars:** For each class interval, draw a rectangle with a width equal to the interval's width and a height proportional to its frequency.

(iv). **Label the Graph:** Clearly label the axes and add an informative title to the histogram.

The resulting histogram provides an immediate visual impression of the data's distribution, revealing key characteristics such as its shape, central tendency, and variability. It is a crucial step in data exploration, allowing us to detect patterns, trends, and potential outliers in the dataset.

Why Not Just a Bar Chart?

It is crucial to distinguish a histogram from a bar chart. A bar chart is typically used for categorical data, showing the count or proportion of distinct categories. The order of bars in a bar chart can often be rearranged without losing meaning. In contrast, a histogram is specifically for numerical data. The order of the bins is fixed, representing a continuous range, and the width of the bars (the bin width) is a critical parameter that directly influences the interpretation of the distribution.

Problem 2.11. *Consider Problem 2.7. Draw histogram of weekly expenditure of 30 students and interpret the results.*

Solution: Consider Problem 2.7. The following histogram (See in Figure 2.8) displays the distribution of weekly expenditures for 30 students. The x-axis

DRAFT

represents the expenditure ranges (bins), and the y-axis represents the number of students in each range.

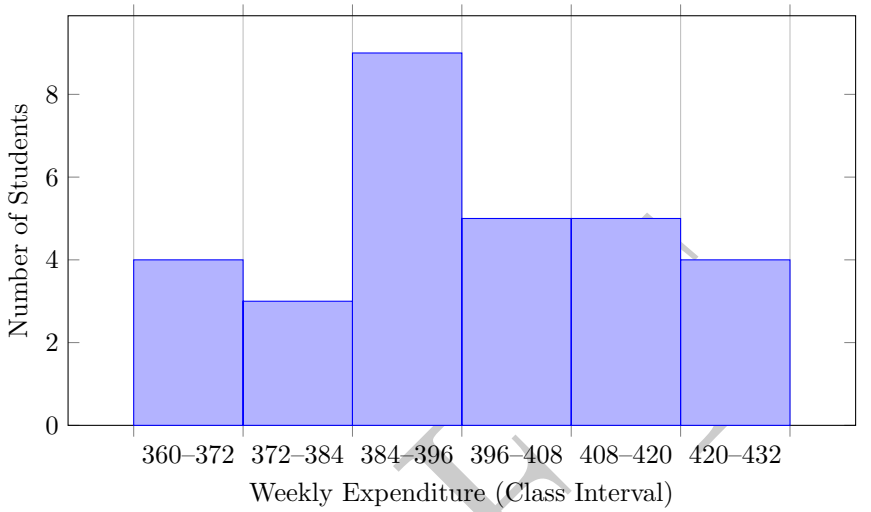


Figure 2.8: Histogram of Weekly Expenditure of 30 Students

Interpreting the Histogram

Firstly, the distribution is *unimodal*, with a clear peak occurring in the class interval **384–396**, where the highest frequency of 9 students is observed. This indicates that most students tend to spend an amount within this range weekly.

The data is spread across the intervals from 360 to 432, with relatively lower frequencies at both ends, specifically in the intervals **360–372** and **420–432**, each containing 4 students. This suggests fewer students spend at the very low or very high ends of the expenditure scale.

The shape of the histogram appears approximately *symmetric*, with frequencies rising gradually to the peak and then decreasing in a similar manner. No significant skewness is observed.

Finally, the histogram does not reveal any apparent outliers, as all class intervals have reasonable frequencies. This visualization effectively summarizes the expenditure pattern of the students, highlighting the central tendency and variability in their weekly spending.

DRAFT

2.3.10 Python Code: Histogram

To create a histogram for the data in Example 2.7 using Python, you can use the following code to generate the histogram.

```
import pandas as pd
import matplotlib.pyplot as plt

# Data: weekly expenditure of 30 students
data = [
423, 369, 387, 411, 393, 394,
371, 377, 389, 409, 392, 407,
431, 401, 363, 391, 405, 382,
400, 381, 399, 415, 428, 422,
395, 371, 410, 419, 386, 390
]

# Convert the data into a pandas DataFrame
df = pd.DataFrame(data, columns=['Expenditure'])

# Define the bin intervals with an interval of 12
bin_start = 360
bin_end = 440
bin_interval = 12
bins = list(range(bin_start, bin_end + bin_interval,
    bin_interval))

# Plot the histogram
plt.figure(figsize=(12, 6))
plt.hist(df['Expenditure'], bins=bins, edgecolor='black',
    alpha=0.7)
plt.xlabel('Weekly Expenditure (Class Interval)')
plt.ylabel('Frequency')
plt.title('Histogram of Weekly Expenditure with Interval 12'
    )
plt.xticks(bins, labels=[f'{edges}-{edges+12-1}' for edges
    in bins[:-1]] + [f'{bins[-1]}+'])
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

2.3.11 Stem-and-Leaf Plot

A **stem-and-leaf plot** is a simple yet powerful way to organize quantitative data to retain the original data values while providing a visual distribution overview. It is particularly useful for small to moderately sized datasets. This plot displays the shape of a distribution and highlights individual data points, which makes it useful for quickly identifying patterns, such as skewness, clus-

ters, gaps, or outliers.

Structure:

- Each number in the dataset is split into two parts:
  - **Stem:** The leading digit(s) of the number.
  - **Leaf:** The final digit.
- The stems are listed in a vertical column in ascending order.
- Each leaf is placed next to its corresponding stem.

Problem 2.12. *The following data represent the test scores of 11 students:*

$$67,\ 72,\ 75,\ 75,\ 78,\ 81,\ 84,\ 85,\ 88,\ 90,\ 91$$

*Construct a **stem-and-leaf plot** for the given data.*

Solution: The corresponding stem-and-leaf plot is:

Stem	Leaves
6	7
7	2 5 5 8
8	1 4 5 8
9	0 1

In this plot:

- The stem "6" represents the 60s, and the leaf "7" represents the score 67.
- The stem "7" with leaves "2 5 5 8" represents scores 72, 75, 75, and 78.
- The stem "8" represents the 80s, and the stem "9" represents the 90s.
- All data values are preserved in this display.

Problem 2.13. *The following data represent the test scores of 11 students:*

$$6.7,\ 7.2,\ 7.5,\ 7.5,\ 7.8,\ 8.1,\ 8.4,\ 8.5,\ 8.8,\ 9.0,\ 9.1$$

*Construct a **stem-and-leaf plot** for this data, where the digits to the left of the decimal point form the stem and the digits to the right of the decimal point form the leaf.*

DRAFT

Solution: The stem-and-leaf plot for the given data is:

Stem	Leaf
6	7
7	2 5 5 8
8	1 4 5 8
9	0 1

Explanation:

- The stem "6" and the leaf "7" represents the data point 6.7.
- The stem "7" with leaves "2 5 5 8" represents the data points 7.2, 7.5, 7.5, and 7.8.
- Similarly, the stem "8" represents the 8.x scores, and the stem "9" represents the 9.x scores.
- All data points are clearly displayed in the plot.

Problem 2.14. *Consider the following dataset representing the blood pressures (in mmHg) of 15 patients.*

$$120, 122, 124, 126, 128, 130, 132, 134, 136, 138, 140, 142, 144, 146, 148$$

Construct a stem-and-leaf plot and comment on it.

Solution: Consider the following dataset representing the systolic blood pressures (in mmHg) of 15 patients:

120, 122, 124, 126, 128, 130, 132, 134, 136, 138, 140, 142, 144, 146, 148

Step 1: Split each number into a stem and a leaf.

- The **stem** consists of the hundreds and tens digit (i.e., the first two digits).
- The **leaf** is the units digit (i.e., the last digit).

Step 2: Construct the stem-and-leaf plot.

Stem	Leaves
12	0 2 4 6 8
13	0 2 4 6 8
14	0 2 4 6 8

DRAFT

Interpretation:

- The data are uniformly distributed across the range of values from 120 to 148 mmHg.
- Each stem has exactly five leaves, indicating a consistent distribution without any clustering or gaps.
- The distribution appears symmetric, with no extreme outliers or skewness.
- The blood pressure readings increase steadily by 2 mmHg, reflecting a regular interval structure in the dataset.

This example illustrates how a stem-and-leaf plot not only preserves all original data values but also provides a clear visual summary of the distribution.

Problem 2.15. ***Construct a stem-and-leaf plot*** *to display the distribution of the following temperatures recorded in Celsius degrees:*

25.32, -3.81, 12, 0.5, -10,
18.861, -7.2, 6, 21.63, -15.4

Create a stem-and-leaf plot appropriately.

Solution: The temperatures recorded in Celsius are:

$$25.32, -3.81, 12, 0.5, -10, 18.861, -7.2, 6, 21.63, -15.4$$

We can construct a stem-and-leaf plot as follows:
Step 1: Round values to one decimal place.

$$25.3, -3.8, 12, 0.5, -10.0, 18.9, -7.2, 6, 21.6, -15.4$$

Step 2: Separate each value into a stem and a leaf.

- The **stem** is the integer part of the number.
- The **leaf** is the first digit after the decimal point.

Step 3: Create the stem-and-leaf plot.

DRAFT

Stem	Leaves
-15	4
-10	0
-7	2
-3	8
0	5
6	0
12	0
18	9
21	6
25	3

Notes:

- Negative stems are listed in increasing order above 0.
- The decimal part of each number is used as the leaf.
- Whole numbers (e.g., 6, 12) are represented with a 0 in the leaf (e.g., 6.0, 12.0).

This stem-and-leaf plot includes both negative and positive temperatures. The distribution is spread out over a range from $-15.4°C$ to $25.3°C$, with no strong clustering. The presence of negative stems indicates colder temperatures, while the positive stems reflect warmer values. This plot retains the original data values and provides a clear visual of how the temperatures are distributed.

DRAFT

2.3.12 Python Code: Stem-and-leaf

To create a stem-and-leaf for the given data in Example 2.15 in Python, we need the following Python code that will generate a stem-and-leaf.

```
from collections import defaultdict

# Data
data = [25, -4, 12, 1, -10, 19, -7, 6, 22, -15]

def stem_and_leaf_plot(data):
# Separate positive and negative values
pos_data = [x for x in data if x >= 0]
neg_data = [-x for x in data if x < 0]

# Determine the stems and leaves
```

```
def create_stem_leaf(data):
stem_leaf = defaultdict(list)
for value in data:
stem = value // 10
leaf = value % 10
stem_leaf[stem].append(leaf)
return stem_leaf

pos_stem_leaf = create_stem_leaf(pos_data)
neg_stem_leaf = create_stem_leaf(neg_data)

# Print positive values
print("Stem-and-Leaf Plot:")
print("Positive values:")
for stem in sorted(pos_stem_leaf.keys()):
leaves = sorted(pos_stem_leaf[stem])
print(f"{stem} | {' '.join(map(str, leaves))}")

# Print negative values
print("Negative values:")
for stem in sorted(neg_stem_leaf.keys()):
leaves = sorted(neg_stem_leaf[stem])
print(f"-{stem} | {' '.join(map(str, leaves))}")

# Generate the stem-and-leaf plot
stem_and_leaf_plot(data)
```

Advantages of Stem-and-Leaf Plot

- Retains original data values.
- Provides a quick visual impression of data distribution.
- For small data sets, stem-and-leaf plots are easy to construct by hand.
- Outliers and gaps in the data can be easily spotted.

Disadvantages of Stem-and-Leaf Plot

- Not effective for very large datasets.
- For data with many decimal places or very large numbers, stem-and-leaf plots may not be practical.
- They are less effective for comparing multiple data sets or groups.

Overall, the stem-and-leaf plot is a valuable exploratory tool that bridges the gap between raw data and formal statistical graphs like histograms.

DRAFT

2.4 Concluding Remarks

As we conclude our exploration of tabular and graphical displays, it is evident that these foundational techniques are indispensable for effective data analysis. The ability to present and interpret data through well-structured tables and informative graphics not only enhances our understanding but also aids in uncovering patterns and trends that might otherwise remain obscured. The skills and Python code examples provided in this chapter equip you with practical tools for summarizing and visualizing data, setting the stage for deeper statistical analysis and more complex data science endeavors. As you move forward, the principles outlined here will serve as a cornerstone for more advanced topics, reinforcing the importance of clear, accurate, and insightful data presentation in the broader field of data science.

2.5 Chapter Exercises

1. You are given the following data from a survey on the incidence of different types of diseases among a sample of patients:

   Flu, Cold, Flu, Allergies, Cold, Flu, Cold, Flu, Allergies, Flu, Flu, Allergies, Cold, Flu, Cold, Flu, Cold, Allergies, Cold, Flu

   Construct a frequency distribution table for the diseases. Draw a bar chart and a pie chart based on the frequency distribution. Interpret the results and comment on the most and least common diseases in the sample.

2. A local community center conducted a survey to find out the preferred recreational activities of its members. The results of the survey are summarized below:

Activity	**Number of Members**
Basketball	45
Swimming	30
Yoga	20
Tennis	15
Running	10

   (a). Calculate the percentage of members who prefer each activity.

   (b). Draw a pie chart to represent the distribution of preferred recreational activities among the members. Write a summary based on this.

DRAFT

3. The following table given in Table 2.8, shows the distribution of sales (in thousands of units) of five different products in a company during the first quarter of the year.

Table 2.8: Sales Distribution of Products

Product	**Sales (thousands of units)**
Laptops	30
Smartphones	20
Tablets	15
Desktops	25
Accessories	10

(a). Calculate the percentage share of each product in the total sales.

(b). Make a pie chart. Write a summary based on this.

4. Imagine you conducted a survey to find out how people spend their leisure time on a typical weekend. You collected data from 100 respondents, and the results are as follows:

- Watching TV: 30 respondents
- Reading: 20 respondents
- Playing Sports: 15 respondents
- Socializing with Friends: 10 respondents
- Playing Video Games: 10 respondents
- Hiking and Outdoor Activities: 8 respondents
- Cooking and Baking: 7 respondents

Answer the following questions:

(a). What is the name of the variable under study? Is this a qualitative variable? If the answer is no, why?

(b). Make a ie chart to represent the distribution of leisure activities among the respondents. Write a summary based on this. Be sure to label the chart accurately and show the percentage of respondents for each activity.

5. The following table shows the number of new students enrolled in a statistics course at a university over five semesters.

DRAFT

Semester	Number of Students
Spring 2022	35
Summer 2022	28
Fall 2022	42
Spring 2023	40
Summer 2023	30

Draw a line diagram to represent the enrollment trend over the semesters.

6. You are given the following set of test scores from a group of students in an engineering course:

$$45, 67, 53, 52, 61, 59, 68, 72, 56, 54,$$
$$63, 75, 49, 62, 60, 58, 66, 64, 55, 70$$

Construct a frequency distribution table using an appropriate number of classes and appropriate class intervals for the given test scores.

7. The following data represents the ages of employees in a company working on a new engineering project:

$$102, 98, 105, 110, 95, 107, 101, 99, 103, 106,$$
$$104, 100, 108, 97, 96, 109, 111, 94, 93, 92$$

Create a frequency distribution table using an appropriate number of classes and appropriate class intervals for the given ages.

8. Here are the weekly sales figures (in units) for a pharmaceutical company over a period of 20 weeks:

$$23, 27, 31, 35, 29, 33, 22, 28, 26, 34,$$
$$32, 25, 30, 24, 36, 21, 37, 38, 39, 20$$

Construct a frequency distribution table using an appropriate number of classes and appropriate class intervals for the given sales figures.

9. The following data represents the weights (in kilograms) of a sample of fruits used in a nutritional study:

$$5.2, 6.3, 7.1, 8.4, 5.5, 6.8, 7.6, 8.0, 5.9, 6.1,$$
$$7.3, 8.7, 5.0, 6.5, 7.8, 8.1, 5.7, 6.9, 7.4, 8.5$$

Make a frequency distribution table using an appropriate number of classes and appropriate class intervals for the given weights.

DRAFT

10. You are given the following set of scores from a recent medical examination:

$$82, 91, 85, 87, 89, 95, 88, 92, 84, 90,$$
$$93, 83, 86, 96, 94, 81, 97, 98, 99, 80$$

Create a frequency distribution table using an appropriate number of classes and appropriate class intervals for the given exam scores.

11. Consider the following data set representing the heights (in cm) of 10 plants:

$$\text{Data: } 150, 155, 160, 162, 165, 168, 170, 175, 180, 185$$

(a). Construct a stem-and-leaf plot for the data set.

(b). What is the maximum number of heights of the data?

DRAFT

Chapter 3

Data Exploration: Numerical Measures

DRAFT

3.1 Introduction

In this chapter, we delve into the fundamental concepts of data exploration with a focus on numerical measures. Understanding these measures is crucial for analyzing and interpreting data effectively. We begin by examining various measures of central tendency, which provide insights into the typical value within a dataset. These include the arithmetic mean, harmonic mean, geometric mean, median, and mode, each with its unique properties, advantages, and limitations.

Following the exploration of central tendency, we will address measures of dispersion or variability. These metrics, such as range, variance, and standard deviation, help us understand the spread and variability of data points around the central value. The chapter will also cover measures of distribution shape, including skewness and kurtosis, which describe the asymmetry and peakedness of the data distribution.

Moreover, we will discuss quartiles, percentiles, and deciles, which are instrumental in dividing the data into meaningful segments, and outline methods for detecting outliers. The chapter concludes with an overview of the five-number summary and boxplots, essential tools for summarizing and visualizing data distribution. Python code will be provided throughout to illustrate practical applications of these concepts.

3.2 Measures of Central Tendency

Measures of central tendency are statistical measures that describe the center or representative value of a dataset. The most common measures of central tendency are the

(i). Mean

- Arithmetic mean
- Harmonic mean
- Geometric mean

(ii). Median

(iii). Mode

These measures provide a single value that represents the middle or center of the data distribution and are essential for summarizing large datasets.

3.2.1 Arithmetic Mean

The **arithmetic mean**, often referred to as the **average**, is the most common measure of central tendency. It is a useful measure when the data is symmetrically distributed without extreme outliers. It is calculated by summing all values and dividing by the number of values.

The arithmetic mean of a set of numbers $x_1, x_2, \ldots, x_n$ is typically denoted by $\bar{x}$ of n and is defined by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Consider a study measuring the time in days for patients to recover from a specific illness, with recovery times being $4, 5, 6, 7, 8$ days. We can use the arithmetic mean to find the average recovery time.

The arithmetic mean is calculated as follows:

$$\bar{x} = \frac{4 + 5 + 6 + 7 + 8}{5} = \frac{30}{5} = 6$$

So, the arithmetic mean of the recovery times is 6 days, which represents the average recovery time.

DRAFT

Properties of Arithmetic Mean

(i). Uniqueness:
For any given data set, the arithmetic mean is unique. There is exactly one value that represents the mean of the data.

(ii). Sum of Deviations is Zero:
The sum of the deviations of all observations from their arithmetic mean is always zero. That is, if $x_1, x_2, \ldots, x_n$ are the observations and $\bar{x}$ is their mean, then

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

(iii). Every Observation Affects the Mean:
Each data value contributes to the calculation of the mean. Thus, changing any one observation will change the value of the mean.

(iv). Mean Lies Between the Extremes:
The arithmetic mean always lies between the minimum and maximum values of the data set. Formally,

$$\min(x_i) \leq \bar{x} \leq \max(x_i).$$

(v). Mean is Sensitive to Outliers:
The arithmetic mean is influenced by extremely large or small values (known as outliers), which may distort the true central tendency of the data.

(vi). Mean of Combined Data:
When two groups of data are combined, the mean of the combined data can be calculated from the means and sizes of the individual groups. If the first group has n_1 observations with mean $\bar{x}_1$ and the second has n_2 observations with mean $\bar{x}_2$, then the combined mean $\bar{x}_c$ is

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$

(vii). Effect of Adding or Subtracting a Constant:
If a constant c is added to (or subtracted from) each observation, the mean changes by the same constant:

$$\bar{x}_{\text{new}} = \bar{x} \pm c.$$

(viii). Effect of Multiplying or Dividing by a Constant:
If each observation is multiplied (or divided) by a constant k, then the mean is also multiplied (or divided) by the same constant:

$$\bar{x}_{\text{new}} = k\bar{x} \quad \text{or} \quad \bar{x}_{\text{new}} = \frac{\bar{x}}{k}.$$

DRAFT

These properties make the arithmetic mean a fundamental and versatile measure for summarizing data.

3.2.2 Advantages and Disadvantages of Arithmetic Mean

The arithmetic mean (or simply the mean) is one of the most commonly used measures of central tendency. It has both advantages and disadvantages depending on the context in which it is used. Understanding its advantages and disadvantages is crucial for selecting the appropriate measure of central tendency for different types of data and analyses.

Advantages

(i) Rigidity and Simplicity: It is rigidly defined, simple, easy to understand, and easy to calculate.

(ii) Uses all data points: It is based upon all the observations in the data set.

(iii) Uniqueness: Its value being unique allows for comparisons between different sets of data.

(iv) Mathematical Properties: The arithmetic mean has useful mathematical properties. For instance, it can be used in further statistical analysis, like calculating variance and standard deviation.

(v) Best for Symmetric Distributions: The arithmetic mean is a reliable measure of central tendency when the data follows a symmetric distribution (like a normal distribution), where the mean is the most representative value.

(vi) Stability: It is least affected by sampling fluctuations compared to other measures of central tendency.

Disadvantages

(i) Sensitivity to Outliers: The mean is highly affected by extreme values or outliers. A few very high or low numbers can skew the mean, making it not represent the "typical" value of the data set.

(ii) Not Suitable for Skewed Distributions: In data sets that are heavily skewed, the mean may not reflect the central location accurately, as it can be pulled toward the tail of the distribution.

(iii) Not Ideal for Non-Numeric Data: The arithmetic mean cannot be applied to nominal or ordinal data, as it requires numerical values to make sense.

DRAFT

(iv) Requirement of Complete Data: It cannot be obtained if a single observation is missing.

(v) Inapplicability with Open Classes: It cannot be calculated if the extreme class is open (e.g., below 10 or above 90).

Problem 3.1. *Suppose we have the following data on the systolic blood pressure (in mmHg) of 10 patients:*

$$120, 130, 125, 140, 135, 128, 132, 138, 124, 126$$

What is the average systolic blood pressure (in mmHg) of 10 patients?

Solution

To calculate the mean systolic blood pressure:

$$\begin{aligned}\bar{x} &= \frac{120 + 130 + 125 + 140 + 135 + 128 + 132 + 138 + 124 + 126}{10} \\ &= \frac{1298}{10} = 129.8 \text{ mmHg}\end{aligned}$$

3.2.3 Geometric Mean

The **geometric mean** is a measure of central tendency that is useful for datasets with exponential growth or multiplicative effects. It is defined as the n-th root of the product of n values.

The geometric mean, denoted by $\bar{x}_{\text{GM}}$, of n values $x_1, x_2, \ldots, x_n$, is defined as

$$\bar{x}_{\text{GM}} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Using the recovery times $4, 5, 6, 7, 8$ days, we calculate the geometric mean to find the average growth rate of recovery.

The geometric mean is calculated as follows:

$$\bar{x}_{\text{GM}} = \sqrt[5]{4 \times 5 \times 6 \times 7 \times 8} = \sqrt[5]{6720} \approx 5.83$$

So, the geometric mean of the recovery times is approximately 5.83 days, reflecting the average multiplicative rate of recovery over time.

DRAFT

Advantages and Disadvantages of Geometric Mean

Advantages

(i) Appropriate for Multiplicative Processes: The geometric mean is ideal for data that involves multiplicative processes, such as rates of growth, financial returns, or other situations where values are multiplied.

(ii) Mitigates the Effect of Extreme Values: It reduces the impact of extremely high or low values in the dataset, which can provide a more balanced measure when dealing with skewed data.

(iii) Handles Proportional Relationships: It is useful for datasets where values are in proportional or percentage terms, such as in economic or financial analysis.

(iv) Stability in Long-Term Growth Rates: In contexts like investment returns, the geometric mean offers a more accurate measure of average growth rates over time compared to the arithmetic mean.

Disadvantages

(i) Cannot Handle Zero or Negative Values: The geometric mean is undefined for datasets containing zero or negative values, as it involves taking the nth root of the product of values.

(ii) Less Intuitive: It is less intuitive and harder to understand compared to the arithmetic mean, making it less accessible for some audiences.

(iii) Requires Logarithmic Transformation: Calculating the geometric mean involves the logarithm of values, which adds complexity compared to simpler averages.

(iv) Sensitive to Variability: While it reduces the impact of extreme values, it may still be affected by large variability in the dataset, especially if values differ greatly.

(v) Not Suitable for All Data Types: It is not appropriate for all types of data, particularly where the data do not naturally fit a multiplicative model or where additive relationships are more relevant.

3.2.4 Harmonic Mean

The **harmonic mean** is a type of average used for rates and ratios. It is defined as the reciprocal of the arithmetic mean of the reciprocals of a set of values.

The harmonic mean of a set of n values $x_1, x_2, \ldots, x_n$, typically denoted

DRAFT

by $\bar{x}_{\text{HM}}$, is defined by

$$\bar{x}_{\text{HM}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Using the recovery times $4, 5, 6, 7, 8$ days, we calculate the harmonic mean to find the average rate of recovery.

The harmonic mean is calculated as follows:

$$\begin{aligned}\bar{x}_{\text{HM}} &= \frac{5}{\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}} \\ &= \frac{5}{0.25 + 0.20 + 0.1667 + 0.1429 + 0.125} \approx 5.65\end{aligned}$$

So, the harmonic mean of the recovery times is approximately 5.65 days, which represents an average recovery time weighted by the rates of recovery.

Advantages and Disadvantages of Harmonic Mean

Advantages

(i) Appropriate for Rates and Ratios: The harmonic mean is particularly useful for averaging rates or ratios, such as speeds, densities, or other quantities where the reciprocal of the average is meaningful.

(ii) Minimizes the Impact of Large Values: It tends to minimize the impact of large values compared to the arithmetic mean. This is beneficial when large values could distort the overall average.

(iii) Emphasizes Small Values: It gives more weight to smaller values in a data set, which can be useful in situations where lower values are more significant or indicative.

(iv) Mathematically Robust: It is less affected by extreme values than the arithmetic mean in datasets where the values are rates or ratios.

Disadvantages

(i) Sensitivity to Zero Values: The harmonic mean cannot be computed if any value in the dataset is zero, as it involves division by the values.

(ii) Less Intuitive: It is less intuitive than the arithmetic mean and is not as commonly used, which can make interpretation and communication more challenging.

DRAFT

(iii) Not Suitable for All Data Types: It is not suitable for data that does not represent rates or ratios. It is not typically used for general numerical data where other means are more appropriate.

(iv) Potential for Misleading Results: In cases where there is significant variability in the data, particularly if large values are present, the harmonic mean can provide misleading results.

(v) Complex Calculation: The calculation of the harmonic mean is more complex compared to the arithmetic mean, which can be a drawback in some practical applications.

3.2.5 Relationships Between Arithmetic Mean, Geometric Mean, and Harmonic Mean

Let $x_1, x_2, \ldots, x_n$ be positive real numbers. The Theorem 3.1 state that the means of a set of positive numbers satisfy the following inequality

$$\bar{x} \geq \bar{x}_{\text{GM}} \geq \bar{x}_{\text{HM}},$$

where $\bar{x}$, $\bar{x}_{\text{GM}}$, and $\bar{x}_{\text{HM}}$ represent the arithmetic mean, geometric mean, and harmonic mean, respectively. This inequality reflects a fundamental property of these measures of central tendency.

Specifically, the arithmetic mean is always greater than or equal to the geometric mean, which in turn is greater than or equal to the harmonic mean. Importantly, equality holds in these inequalities if and only if all the elements in the dataset are equal. In other words, $\bar{x} = \bar{x}_{\text{GM}} = \bar{x}_{\text{HM}}$ if and only if $x_1 = x_2 = \cdots = x_n$.

Theorem 3.1. *Let $x_1, x_2, \ldots, x_n$ be positive real numbers. Let $\bar{x}_{AM}$ denote the arithmetic mean, $\bar{x}_{GM}$ denote the geometric mean, and $\bar{x}_{HM}$ denote the harmonic mean of these numbers. Then,*

$$\bar{x}_{AM} \geq \bar{x}_{GM} \geq \bar{x}_{HM}$$

Equality holds in both inequalities if and only if $x_1 = x_2 = \cdots = x_n$.

Proof. We will prove the inequality in two parts.

Part 1: Proof of $\bar{x}_{AM} \geq \bar{x}_{GM}$ (AM-GM Inequality)

We will use the property that for a convex function f, Jensen's inequality holds:

$$f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Consider the function $f(x) = -\ln(x)$ for $x > 0$. The first derivative is $f'(x) = -\frac{1}{x}$, and the second derivative is $f''(x) = \frac{1}{x^2}$. Since $f''(x) > 0$ for all

DRAFT

$x > 0$, the function $f(x) = -\ln(x)$ is strictly convex on the interval $(0, \infty)$.

Applying Jensen's inequality to $f(x) = -\ln(x)$ and the positive numbers $x_1, x_2, \ldots, x_n$:

$$-\ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n}(-\ln(x_i))$$

$$-\ln\left(\bar{x}_{AM}\right) \leq -\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$$

Using the property of logarithms $\sum_{i=1}^{n}\ln(x_i) = \ln\left(\prod_{i=1}^{n} x_i\right)$:

$$-\ln\left(\bar{x}_{AM}\right) \leq -\frac{1}{n}\ln\left(\prod_{i=1}^{n} x_i\right)$$

$$-\ln\left(\bar{x}_{AM}\right) \leq -\ln\left(\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right)$$

$$-\ln\left(\bar{x}_{AM}\right) \leq -\ln\left(\bar{x}_{GM}\right)$$

Since the natural logarithm function $\ln(x)$ is strictly increasing, the function $-\ln(x)$ is strictly decreasing. Therefore, multiplying by -1 and reversing the inequality sign gives:

$$\ln\left(\bar{x}_{AM}\right) \geq \ln\left(\bar{x}_{GM}\right).$$

Again, due to the strictly increasing nature of $\ln(x)$, we conclude:

$$\bar{x}_{AM} \geq \bar{x}_{GM}$$

Equality in Jensen's inequality for a strictly convex function holds if and only if all the arguments are equal, i.e., $x_1 = x_2 = \cdots = x_n$. Therefore, equality in $\bar{x}_{AM} \geq \bar{x}_{GM}$ holds if and only if $x_1 = x_2 = \cdots = x_n$.

Part 2: Proof of $\bar{x}_{GM} \geq \bar{x}_{HM}$ (GM-HM Inequality)

Consider the reciprocals of the positive numbers $x_1, x_2, \ldots, x_n$, which are $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$. Applying the AM-GM inequality (proven in Part 1) to these positive numbers, we have:

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{n} \geq \left(\frac{1}{x_1}\cdot\frac{1}{x_2}\cdot\cdots\cdot\frac{1}{x_n}\right)^{1/n}$$

$$\frac{\sum_{i=1}^{n}\frac{1}{x_i}}{n} \geq \left(\frac{1}{\prod_{i=1}^{n} x_i}\right)^{1/n}$$

$$\frac{\sum_{i=1}^{n}\frac{1}{x_i}}{n} \geq \frac{1}{\left(\prod_{i=1}^{n} x_i\right)^{1/n}}$$

DRAFT

$$\frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n} \geq \frac{1}{\bar{x}_{GM}}$$

Taking the reciprocal of both sides of the inequality. Since both sides are positive, the inequality sign reverses:

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \leq \bar{x}_{GM}$$

By definition, the left side is the harmonic mean $\bar{x}_{HM}$:

$$\bar{x}_{HM} \leq \bar{x}_{GM}$$

Which is equivalent to:

$$\bar{x}_{GM} \geq \bar{x}_{HM}$$

The equality in the AM-GM inequality applied to $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$ holds if and only if $\frac{1}{x_1} = \frac{1}{x_2} = \cdots = \frac{1}{x_n}$. This condition is equivalent to $x_1 = x_2 = \cdots = x_n$. Therefore, equality in $\bar{x}_{GM} \geq \bar{x}_{HM}$ holds if and only if $x_1 = x_2 = \cdots = x_n$.

Combining **Part 1** and **Part 2**, we have proven the AM-GM-HM inequality:

$$\bar{x}_{AM} \geq \bar{x}_{GM} \geq \bar{x}_{HM}$$

with equality holding throughout if and only if $x_1 = x_2 = \cdots = x_n$. □

When to Use Arithmetic, Geometric, and Harmonic Means

The three commonly used means are Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM). Each has a specific purpose and is suitable for different types of data and contexts.

- **Arithmetic Mean (AM):**
  - Best suited for data measured on the same scale and where values are additive.
  - Commonly used for typical average calculations, such as mean test scores, average temperatures, or average income.
- **Geometric Mean (GM):**
  - Appropriate when dealing with data that involves products, growth rates, or ratios.
  - Commonly used to calculate average rates of change over time, such as compound interest or population growth rates.
- **Harmonic Mean (HM):**

DRAFT

- Suitable for data expressed as rates or ratios, especially when values are in reciprocal relationships.
- Commonly used for averaging speeds, price per unit, or rates of consumption.

Choosing the appropriate mean is essential for accurate analysis and interpretation of data in various applications.

3.2.6 Median

The median is a measure of central tendency that divides a dataset into two equal halves. Let $x_1, x_2, \ldots, x_n$ be a set of n numerical observations. To find the median, first arrange the data in ascending (or descending) order. Let the ordered data set be denoted by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(i)}$ is the i-th value in the ordered set.

The median is then defined as follows:
Let $m = \frac{n+1}{2}$.

$$\text{Median} = \begin{cases} x_{(m)} & \text{if } m \text{ is an integer (i.e., } n \text{ is odd)} \\ \frac{x_{(\lfloor m \rfloor)} + x_{(\lceil m \rceil)}}{2} & \text{if } m \text{ is not an integer (i.e., } n \text{ is even)} \end{cases}$$

Here, $\lfloor m \rfloor$ is the floor function (the greatest integer less than or equal to m), and $\lceil m \rceil$ is the ceiling function (the smallest integer greater than or equal to m).

When there is an **odd number** of observations, the median is simply the middle value. For example, consider a dataset of seven blood pressure readings:

$$110, 115, 120, 125, 130, 135, 140.$$

Since there are seven observations (an odd number), the median is the fourth value. In this case, $m = \frac{n+1}{2} = \frac{7+1}{2} = 4$ which is an integer and hence, the median is

$$\text{Median} = x_{(4)} = 125.$$

Conversely, when there is an **even** number of observations, there is no single middle value, so the median is calculated as the average of the two middle values. For instance, in a dataset of eight cholesterol levels:

$$200, 210, 220, 225, 230, 240, 250, 260,$$

the median is found by taking the average of the fourth and fifth values. In this

DRAFT

case, $m = \frac{n+1}{2} = \frac{8+1}{2} = 4.5$ is not an integer. Hence, the median is

$$\begin{aligned}\text{Median} &= \frac{x_{(\lfloor 4.5 \rfloor)} + x_{(\lceil 4.5 \rceil)}}{2} \\ &= \frac{x_{(4)} + x_{(5)}}{2} = \frac{225 + 230}{2} = 227.5\end{aligned}$$

Median is particularly useful in data science for understanding the typical value in a dataset that is not skewed by extreme values.

Problem 3.2. *Consider the following data on the number of hours of sleep per night for a group of 9 adults:*

$$7, 8, 5, 6, 9, 7, 6, 10, 8$$

What is the median hours of sleep per night?

Solution

First, arrange the data in ascending order:

$$5, 6, 6, 7, 7, 8, 8, 9, 10$$

In this case, $n = 9$, so $m = \frac{n+1}{2} = \frac{9+1}{2} = 5$, which is an integer. Hence,

$$\text{Median} = x_{(m)} = x_{(5)} = 7 \text{ hours}$$

□

Problem 3.3. *Consider the following data on the test scores of 6 students:*

$$85, 92, 78, 95, 88, 80$$

What is the median test score?

Solution

First, arrange the data in ascending order:

$$78, 80, 85, 88, 92, 95$$

In this case, $n = 6$, so $m = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$, which is not an integer. Hence, we use the second case of the median formula,

$$\text{Median} = \frac{x_{(\lfloor m \rfloor)} + x_{(\lceil m \rceil)}}{2} = \frac{x_{(\lfloor 3.5 \rfloor)} + x_{(\lceil 3.5 \rceil)}}{2} = \frac{x_{(3)} + x_{(4)}}{2}$$

From the ordered data, $x_{(3)} = 85$ and $x_{(4)} = 88$. Therefore,

$$\text{Median} = \frac{85 + 88}{2} = \frac{173}{2} = 86.5$$

The median test score is 86.5. □

DRAFT

3.2.7 Advantages and Disadvantages of Median

Advantages of Median

(i) Robust to Outliers: The median is not affected by extreme values or outliers, providing a better central measure for skewed distributions.

(ii) Simple to Compute: It is easy to calculate, especially for small datasets, by finding the middle value when data is ordered.

(iii) Represents the 50th Percentile: It divides the dataset into two equal halves, making it a useful measure for understanding the data's center.

(iv) Applicable to Ordinal Data: Can be used with ordinal data, where data values are ranked but not necessarily numeric.

Disadvantages of Median

(i) Ignores Data Values: Does not take into account the magnitude of all data values, only their order.

(ii) Less Informative for Symmetric Distributions: Provides less information about the distribution of data than the mean, especially if the data is symmetric.

(iii) Not Suitable for Further Mathematical Operations: Unlike the mean, the median cannot be easily used in further mathematical or statistical calculations.

(iv) Not Unique for Small Datasets: In small datasets with an even number of values, there may be two middle values, which can complicate interpretation.

3.2.8 Mode

The mode is the value that appears most frequently in a dataset. A dataset can have more than one mode if multiple values have the same highest frequency. The mode is useful for categorical data or when identifying the most common value.

Problem 3.4. *Consider a manufacturing process where engineers are measuring the diameter of a set of machine components to ensure they meet quality specifications. The following is a list of diameters (in millimeters) of 20 components that were measured:*

$$50, 52, 51, 50, 53, 52, 54, 50, 51, 52, 55, 50, 53, 50, 51, 52, 55, 50, 52, 51$$

Find the mode of this dataset.

DRAFT

Solution

To find the mode of this dataset:

$$50, 52, 51, 50, 53, 52, 54, 50, 51, 52, 55, 50, 53, 50, 51, 52, 55, 50, 52, 51$$

Count the frequency of each diameter:

- Diameter 50 occurs 6 times.
- Diameter 52 occurs 5 times.
- Diameter 51 occurs 4 times.
- Diameters 53 and 55 occur 2 times each.
- Diameter 54 occurs 1 time.

We see the diameter 50 appears most frequently (6 times). Thus, the mode of this dataset is 50 mm.

Problem 3.5. *Suppose we have the blood groups of 15 patients recorded as follows:*

$$A+, B+, O+, A+, B+, O+, A+, B\text{-}, A+, A\text{-}, B+, O\text{-}, A+, O+, B+$$

What is the mode of the blood groups?

Solution

The frequency distribution of blood types is:

- Blood group A+: 5 patients
- Blood group A-: 1 patient
- Blood group B+: 4 patients
- Blood group B-: 1 patient
- Blood group O+: 3 patients
- Blood group O-: 1 patient

The blood group with the highest frequency is **A+**, occurring 5 times. Thus, the mode of the dataset is blood group A+.

DRAFT

3.2.9 Advantages and Disadvantages of Mode

Advantages of Mode

(i) Easy to Identify: The mode is straightforward to determine as it is simply the most frequently occurring value in a dataset.

(ii) Applicable to All Data Types: Can be used with nominal, ordinal, and some quantitative data, making it versatile.

(iii) Reflects Commonality: Represents the value or values that occur most often, which can be useful for understanding common trends or preferences.

(iv) Handles Categorical Data: Ideal for categorical data where numerical calculations are not applicable.

Disadvantages of Mode

(i) May Not Be Unique: A dataset can have more than one mode or no mode at all, which can complicate interpretation.

(ii) Not Useful for Continuous Data: Less useful for continuous data with many unique values, as identifying the most frequent value can be challenging.

(iii) Does Not Reflect Data Distribution: Does not provide information about the spread or shape of the data distribution.

(iv) Insensitive to Changes: The mode does not account for changes in data that do not affect frequency, potentially overlooking variations.

3.2.10 Choosing the Ideal Measure of Central Tendency

The choice of the ideal measure of central tendency depends on the nature of the data and the analysis objectives. The primary measures are the arithmetic mean, median, and mode. Here is a guide to selecting the most appropriate measure:

- **Arithmetic Mean:** Best for symmetric distributions and numerical data.
- **Median:** Best for skewed distributions and when robustness to outliers is needed.
- **Mode:** Best for categorical data and identifying the most frequent values.

DRAFT

3.2.11 Weighted Mean

The **weighted mean** is a type of average that accounts for the relative importance or frequency of each data value. Unlike the simple arithmetic mean, which treats each observation equally, the weighted mean assigns a specific weight to each value, reflecting its contribution to the overall average.

Weighted Mean: For a set of n observations $x_1, x_2, \ldots, x_n$ with corresponding weights $w_1, w_2, \ldots, w_n$, the weighted mean $\bar{x}_w$ is defined as:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where,

- x_i is the i^{th} observation,
- w_i is the weight assigned to x_i,
- $\sum_{i=1}^{n} w_i$ is the sum of the weights.

Problem 3.6. *A biostatistician is analyzing the average blood pressure readings of three different age groups in a study. The average blood pressure for each age group and the number of individuals in each group are as follows:*

- *Age Group 1: Average blood pressure* $= 120$ *mmHg, Number of individuals* $= 30$
- *Age Group 2: Average blood pressure* $= 130$ *mmHg, Number of individuals* $= 25$
- *Age Group 3: Average blood pressure* $= 140$ *mmHg, Number of individuals* $= 20$

Calculate the weighted mean of the average blood pressure across all age groups.

Solution:

The weighted mean is

$$\bar{x}_w = \frac{(120 \times 30) + (130 \times 25) + (140 \times 20)}{30 + 25 + 20}$$

$$\bar{x}_w = \frac{3600 + 3250 + 2800}{75} = \frac{9650}{75} \approx 128.67 \text{ mmHg}$$

Therefore, the weighted mean blood pressure is approximately 128.67 mmHg.

Problem 3.7. *A clinical trial evaluates the efficacy of a new drug. The efficacy rates are measured as follows:*

DRAFT

- *Trial 1: Efficacy* $= 85\%$, *Weight* $= 10$
- *Trial 2: Efficacy* $= 90\%$, *Weight* $= 20$
- *Trial 3: Efficacy* $= 80\%$, *Weight* $= 15$

Find the weighted mean efficacy rate of the drug.

Solution:

The weighted mean is

$$\begin{aligned}\bar{x}_w &= \frac{(85 \times 10) + (90 \times 20) + (80 \times 15)}{10 + 20 + 15} \\ &= \frac{850 + 1800 + 1200}{45} \\ &= \frac{3850}{45} \approx 85.56\%\end{aligned}$$

Hence, the weighted mean efficacy rate is approximately 85.56%.

Problem 3.8. *A public health researcher records the incidence rates of a disease in three different regions. The incidence rates and the population sizes for each region are:*

- *Region A: Incidence rate* $= 0.02$ *cases per person, Population* $= 50,000$
- *Region B: Incidence rate* $= 0.03$ *cases per person, Population* $= 75,000$
- *Region C: Incidence rate* $= 0.04$ *cases per person, Population* $= 25,000$

Solution:

The weighted mean is

$$\begin{aligned}\bar{x}_w &= \frac{(0.02 \times 50{,}000) + (0.03 \times 75{,}000) + (0.04 \times 25{,}000)}{50{,}000 + 75{,}000 + 25{,}000} \\ &= \frac{1000 + 2250 + 1000}{150{,}000} \\ &= \frac{4250}{150{,}000} \approx 0.02833 \text{ cases per person}\end{aligned}$$

Hence, the weighted mean incidence rate is approximately 0.02833 cases per person.

Problem 3.9. *An academic advisor evaluates the performance of students based on their grades in three courses, with different credit hours for each course:*

- *Course 1: Grade* $= 85$, *Credits* $= 3$
- *Course 2: Grade* $= 90$, *Credits* $= 4$
- *Course 3: Grade* $= 80$, *Credits* $= 2$

Compute the weighted mean grade.

DRAFT

Solution:

The weighted mean is

$$\begin{aligned}\bar{x}_w &= \frac{(85 \times 3) + (90 \times 4) + (80 \times 2)}{3 + 4 + 2} \\ &= \frac{255 + 360 + 160}{9} \\ &= \frac{775}{9} \approx 86.11\end{aligned}$$

Hence, theweighted mean grade is approximately 86.11.

3.2.12 Measures of Central Tendency for Grouped Data

In this section, we demonstrate how to estimate the mean, median and mode of grouped data using the example of weekly expenditure of 30 students give in Problem 2.7. Suppose we have the weekly expenditure of 30 students, and the frequency distribution of this weekly expenditure is presented in Table 3.1. The mean, median, and mode can be calculated in the following ways:

Table 3.1: Distribution of Weekly Expenditure of 30 Students

Class Interval	**Frequency** (f_i)
360 - 372	4
372 - 384	3
384 - 396	9
396 - 408	5
408 - 420	5
420 - 432	4
Total	30

DRAFT

Mean Estimation

The mean (arithmetic mean) for grouped data is given by the formula:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

where f_i is the frequency and x_i is the midpoint of each class interval.

Example: For frequency table given in Table 3.1, we calculate the midpoint (class mark) x_i for each class interval using the following formula:

$$x_i = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

The midpoints are presented in Table 3.2. We now compute the products $f_i x_i$ and presented in the fourth column in Table 3.1.

Table 3.2: Calculation of $f_i x_i$

Class Interval	Frequency (f_i)	Midpoint(x_i)	$f_i x_i$
360 – 372	4	366	1464
372 – 384	3	378	1134
384 – 396	9	390	3510
396 – 408	5	402	2010
408 – 420	5	414	2070
420 – 432	4	426	1704
Total	30		11892

Hence, the group mean is

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{11892}{30} = 396.4$$

Thus, the estimated mean weekly expenditure for the 30 students is **approximately 396.4**. This mean value provides a central tendency estimate for the grouped data.

Median Estimation

The median for grouped data is given by the formula:

$$\text{Median} = l + \left(\frac{\frac{n}{2} - cf}{f} \right) \times h$$

where l is the lower boundary of the median class, n is the total frequency, cf is the cumulative frequency before the median class, f is the frequency of the **median class**, and h is the class width.

Example: The frequency distribution with the cumulative frequency (cf) is presented in Table 3.3.

DRAFT

Table 3.3: Weekly expenditure data with cumulative frequency

Class Interval	**Frequency** (f_i)	**Cumulative Frequency** (cf_i)
360 – 372	4	4
372 – 384	3	7
384 – 396	9	16
396 – 408	5	21
408 – 420	5	26
420 – 432	4	30
Total	30	

For finding **median class**, the median is at the $\left(\frac{n}{2}\right)$th observation:

$$\frac{n}{2} = \frac{30}{2} = 15$$

The cumulative frequency just before 15 is 7, so the **median class** is 384 – 396 and

$$l = 384, \quad n = 30, \quad cf = 7, \quad f = 9, \quad h = 12$$

$$\text{Median} = 384 + \left(\frac{\frac{30}{2} - 7}{9}\right) \times 12 = 384 + \left(\frac{15 - 7}{9}\right) \times 12$$
$$= 384 + \left(\frac{8}{9}\right) \times 12 = 384 + 10.67 = 394.67$$

Mode Estimation

The mode for grouped data is given by the formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)}\right) \times h$$

where l is the lower boundary of the modal class, f_1 is the frequency of the modal class, f_0 is the frequency of the class before the modal class, f_2 is the frequency of the class after the modal class, and h is the class width.

$$\text{Modal class: } 384 - 396$$

$$l = 384, \quad f_1 = 9, \quad f_0 = 3, \quad f_2 = 5, \quad h = 12$$

DRAFT

$$\text{Mode} = 384 + \left(\frac{9-3}{(9-3)+(9-5)}\right) \times 12 = 384 + \left(\frac{6}{6+4}\right) \times 12$$
$$= 384 + \left(\frac{6}{10}\right) \times 12 = 384 + 7.2 = 391.2$$

3.2.13 Python Code: Mean, Median and Mode

In this section, we will demonstrate how to compute the arithmetic mean, geometric mean, harmonic mean, median, and mode using Python. Consider the following dataset:

$$\{10, 15, 15, 20, 25, 30\}$$

```python
import numpy as np
from scipy.stats import gmean, hmean, mode

# Sample data
data = [10, 15, 15, 20, 25, 30]

# Arithmetic Mean
arithmetic_mean = np.mean(data)
print(f"Arithmetic Mean: {arithmetic_mean}")

# Geometric Mean
geometric_mean = gmean(data)
print(f"Geometric Mean: {geometric_mean}")

# Harmonic Mean
harmonic_mean = hmean(data)
print(f"Harmonic Mean: {harmonic_mean}")

# Median
median = np.median(data)
print(f"Median: {median}")

# Mode
mode_value, count = mode(data)
print(f"Mode: {mode_value}")
```

The output of the code will be:

- **Arithmetic Mean:** 19.1667
- **Geometric Mean:** 17.9768
- **Harmonic Mean:** 16.8224
- **Median:** 17.5
- **Mode:** 15

3.2.14 Exercises

1. Suppose we have the following dataset representing the scores of students in a test:
$$85, 90, 78, 92, 88, 76, 95, 89, 84, 91$$
   (a) Calculate the arithmetic mean of the test scores.
   (b) Determine the median of the test scores.

2. Consider the following dataset representing the monthly salaries (in thousands of dollars) of 12 employees at a company:
$$45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100$$
   (a) Calculate the arithmetic mean of the salaries.
   (b) Determine the median salary.

3. A company measures the speed of data transmission across four networks as follows (in Mbps):
$$10, 20, 30, 40$$
   (a) Calculate the arithmetic mean of the data transmission speeds.
   (b) Compute the harmonic mean of the data transmission speeds.
   (c) Calculate the median of the data transmission speeds.

4. The following dataset represents the time (in hours) taken by a worker to complete four tasks:
$$4, 6, 8, 12$$
   (a) Calculate the arithmetic mean of the task completion times.
   (b) Compute the harmonic mean of the task completion times.

5. A company tracks the monthly returns (in percentage) on three different investments over a year:
$$4\%, 8\%, 12\%$$
   (a) Calculate the geometric mean of the monthly returns.
   (b) Compute the harmonic mean of the monthly returns.

6. A survey reports the annual growth rates of four investments as follows (in percentage):
$$5\%, 10\%, 15\%, 20\%$$
   (a) Calculate the geometric mean of the annual growth rates.
   (b) Compute the harmonic mean of the annual growth rates.

DRAFT

7. Consider the following dataset of the number of books read by a group of 15 students:

$$2, 3, 2, 5, 7, 2, 4, 5, 5, 3, 6, 7, 8, 2, 5$$

   (a) Identify the mode(s) of the number of books read.

   (b) Calculate the mean of the number of books read.

8. A class of 20 students recorded their scores on a recent quiz as follows:

$$82, 85, 88, 82, 90, 85, 92, 82, 88, 85, 87, 90, 82, 85, 92, 90, 87, 88, 85, 82$$

   (a) Compute the mean of the quiz scores.

   (b) Identify the mode of the quiz scores.

9. The following dataset represents the hours spent on three different activities by a student each week, with respective weights:

   - Activity Hours:
   $$10, 15, 20$$
   - Weights:
   $$2, 3, 1$$

   Compute the weighted mean of the hours spent on activities.

10. A company reports the sales (in units) of three different products with the following weights:

    - Sales:
    $$200, 300, 400$$
    - Weights:
    $$1, 3, 2$$

    Calculate the weighted mean of the sales data.

11. Define the weighted mean. A student's overall grade in a course is determined by three components: homework, exams, and projects. The weights assigned to each component are as follows: homework (30%), exams (50%), and projects (20%). The student scored 85 in homework, 90 in exams, and 80 in projects. Calculate the student's overall weighted mean grade.

12. A dataset is grouped into the following intervals with corresponding frequencies:

    - Class Intervals:
    $$[10 - 20, 20 - 30, 30 - 40, 40 - 50]$$

DRAFT

- Frequencies:

$$[8, 15, 20, 12]$$

(a) Calculate the mean and median of the grouped data.

(b) Calculate the mode of the grouped data.

13. A student receives grade points in three subjects with the following weights:

- Grade points:

$$3.5, 3.7, 4$$

- Credit:

$$3, 4, 2$$

Compute an average of the grade points.

14. A healthcare provider tracks the number of patients visiting a clinic over a week (in patients per day) as follows:

$$12, 15, 14, 17, 15, 18, 16$$

(a) Find the median number of patients per day.

(b) Determine the mode of the number of patients.

3.3 Measures of Dispersion or Variability

Measures of dispersion or variability describe the spread or distribution of data points in a dataset. They describe the extent to which data values differ from the central tendency, such as the mean or median. These measures help to understand how much the data varies around the central tendency (mean, median, or mode).

They provide insights into the distribution's spread and consistency. They are categorized into **absolute** and **relative measures**. Here's an overview of both:

1. Absolute Measures of Dispersion
   - Range
   - Interquartile range (IQR)
   - Semi-interquartile range
   - Variance (or standard deviation)
   - Mean Absolute Deviation (MAD), etc.
2. Relative Measures of Dispersion

DRAFT

- Relative Range
- Coefficient of variation
- Skewness
- Kurtosis, etc.

- **Absolute Measures** give the dispersion in the same units as the data and include range, standard deviation (and variance), and mean absolute deviation.
- **Relative Measures** provide dispersion relative to the mean or other central values and include the coefficient of variation, relative range, and quartile coefficient of dispersion.

We will explain some important measures of dispersion in detail in the following sections.

3.3.1 Range

The range is the simplest measure of variability and is calculated as the difference between the maximum and minimum values in a dataset. Mathematically, for a dataset $\{x_1, x_2, \ldots, x_n\}$, the range is defined as:

$$\text{Range} = x_{\max} - x_{\min}$$

where $x_{\max}$ is the maximum and $x_{\min}$ is the minimum observation in the dataset.

In a study measuring the blood glucose levels of 10 patients, the highest reading is 120 mg/dL and the lowest reading is 85 mg/dL.So, the range is

$$\text{Range} = 120 - 85 = 35 \text{ mg/dL}$$

3.3.2 Variance

Variance measures the average squared deviation of each data point from the mean. It reflects how data points spread out around the mean. For a sample of size n with observations $x_1, x_2, \ldots, x_n$, and sample mean $\bar{x}$, the sample variance s^2 is defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where

- x_i is the i-th observation,

DRAFT

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

Note that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$

so it is often convenient to use the following formula:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

where $\bar{x}$ is the arithmetic mean, x_i represents the data points, and n is the number of data points.

Consider the weights of 5 patients: $65, 70, 75, 80, 85$ kg. We have,

$$\bar{x} = \frac{65 + 70 + 75 + 80 + 85}{5} = 75$$

Weight (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
65	-10	100
70	-5	25
75	0	0
80	5	25
85	10	100
Sum		250

Thus the variance is

$$s^2 = \frac{250}{4} = 62.5 \text{ kg}^2$$

3.3.3 Standard Deviation

The standard deviation is the positive square root of the variance. It provides a measure of dispersion in the same units as the data, making it easier to interpret. Mathematically, the sample standard deviation, denoted by s, is defined as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

A more convenient formula for calculating the standard deviation is:

DRAFT

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}.$$

Using the variance from the previous example, the sample standard deviation is:

$$s = \sqrt{62.5} \approx 7.91 \text{ kg}.$$

Problem 3.10. *A physician collected an initial measurement of hemoglobin (g/L) after the admission of 10 inpatients to a hospital's department of cardiology. The hemoglobin measurements were 139, 158, 120, 112, 122, 132, 97, 104, 159, and 129 g/L. Calculate the variance and standard deviation of hemoglobin level.*

Solution

The variance is calculated by the following way:

x_i	x_i^2
139	19321
158	24964
120	14400
112	12544
122	14884
132	17424
97	9409
104	10816
159	25281
129	16641
$\sum_{i=1}^{n} x_i = 1272$	$\sum_{i=1}^{n} x_i^2 = 165684$

DRAFT

The sample Mean and Variance:

$$\begin{aligned}\bar{x} &= \frac{1272}{10} = 127.2 \text{ g/L} \\ s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\ &= \frac{1}{10-1}\left(165684 - 10 \times 127.2^2\right) \\ &= 431.7 \text{ (g/L)}^2\end{aligned}$$

The sample standard deviation is $s = \sqrt{431.7} = 20.8$ g/L. Thus, the standard deviation of hemoglobin for these 10 patients was 20.8 g/L.

3.3.4 Mean Absolute Deviation (MAD)

The **Mean Absolute Deviation (MAD)** is a measure of dispersion that tells us how much the values in a data set deviate, on average, from the mean of the data set. Unlike the variance and standard deviation, which square the deviations, the MAD simply uses the absolute value of the deviations. This makes it a more intuitive measure of variability in many contexts.

For a data set consisting of n observations $x_1, x_2, \ldots, x_n$, with mean $\bar{x}$, the **Mean Absolute Deviation** is given by:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| .$$

The MAD measures the average distance of each data point from the mean. A smaller MAD indicates that the data points are closer to the mean, showing less variability. A larger MAD suggests more variability in the data. The MAD is **less sensitive** to outliers compared to the standard deviation because it does not involve squaring the deviations. It is especially useful for understanding the **average magnitude** of deviations in a data set, regardless of direction.

Example

Let us consider the following data set of test scores:

$$4, \quad 6, \quad 8, \quad 10, \quad 12.$$

We have,

$$\bar{x} = \frac{4+6+8+10+12}{5} = \frac{40}{5} = 8.$$

DRAFT

Observation (x_i)	**Deviation from Mean** $(x_i - \bar{x})$	**Absolute Deviation** $\lvert(x_i - \bar{x})\rvert$
4	$4 - 8 = -4$	4
6	$6 - 8 = -2$	2
8	$8 - 8 = 0$	0
10	$10 - 8 = 2$	2
12	$12 - 8 = 4$	4
Total	**0**	**12**

So, the MAD is

$$\text{MAD} = \frac{12}{5} = 2.4.$$

Thus, the Mean Absolute Deviation of this data set is 2.5.

Problem 3.11. *A physician collected an initial measurement of hemoglobin (g/L) after the admission of 10 inpatients to a hospital's department of cardiology. The hemoglobin measurements were:*

$$139, \quad 158, \quad 120, \quad 112, \quad 122, \quad 132, \quad 97, \quad 104, \quad 159, \quad 129.$$

Calculate the mean absolute deviation of the hemoglobin level.

Solution

The mean hemoglobin level is:

$$\bar{x} = \frac{139 + 158 + 120 + 112 + 122 + 132 + 97 + 104 + 159 + 129}{10} = 127.2\,\text{g/L}.$$

The absolute deviations from the mean are presented in the following table:

DRAFT

Observation (x_i)	**Deviation** $(x_i - \bar{x})$	**Absolute Deviation** $\lvert x_i - \bar{x}\rvert$
139	11.8	11.8
158	30.8	30.8
120	−7.2	7.2
112	−15.2	15.2
122	−5.2	5.2
132	4.8	4.8
97	−30.2	30.2
104	−23.2	23.2
159	31.8	31.8
129	1.8	1.8
Total		**162.0**

Thus, the mean absolute deviation is

$$\text{MAD} = \frac{162.0}{10} = 16.2\,\text{g/L}.$$

3.3.5 Measures of Variability for Grouped Data

Variance for Grouped Data

Let $x_1, x_2, \ldots, x_k$ be the midpoints of class intervals, and $f_1, f_2, \ldots, f_k$ be the corresponding frequencies. Then, the variance for grouped data is given by the formula:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}$$

where $n = \sum_{i=1}^{k} f_i$.

Standard deviation for Grouped Data

Standard deviation is simply the positive square root of variance. Hence, the sndard deviation is defined as

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1}}$$

To compute the sample variance and standard deviation for grouped data, a detailed calculation with an example is given below.

DRAFT

Example: Consider the weekly expenditure of 30 students given in Problem 2.7. The frequency distribution of this data is shown in Table 3.4. The grouped variance and standard deviation can be calculated using the midpoint method as follows:

Table 3.4: Distribution of Weekly Expenditure of 30 Students

Class Interval	**Frequency** (f_i)
360 – 372	4
372 – 384	3
384 – 396	9
396 – 408	5
408 – 420	5
420 – 432	4
Total	30

Table 3.5: Computation of Sample Variance for Grouped Data

f_i	x_i	$f_i x_i$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
4	366	1464	912.04	3648.16
3	378	1134	331.24	993.72
9	390	3510	38.44	345.96
5	402	2010	33.64	168.20
5	414	2070	316.84	1584.20
4	426	1704	888.04	3552.16
Total		**11892**		**10292.40**

The sample mean is:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{11892}{30} = 396.4$$

The sample variance is:

$$s^2 = \frac{10292.40}{30 - 1} = \frac{10292.40}{29} \approx 354.91$$

And the sample standard deviation is:

$$s = \sqrt{354.91} \approx 18.83$$

DRAFT

3.3.6 Coefficient of Variation

The coefficient of variation (CV) is a normalized measure of dispersion that expresses the standard deviation as a percentage of the mean. It' is often used to compare the degree of variation from one dataset to another, even if their means are drastically different.

Let us consider a sample $x_1, x_2, \ldots, x_n$ of size n. Then, the sample coefficient of variation can be defined as follows:

$$\text{CV} = \frac{s}{\bar{x}} \times 100\%$$

where

- s = sample standard deviation
- $\bar{x}$ = sample mean

If the mean blood pressure is $\bar{x} = 128.67$ mmHg and the standard deviation is $s = 7.91$ mmHg, the CV is:

$$\text{CV} = \frac{7.91}{128.67} \times 100\% \approx 6.15\%$$

Problem 3.12. *A study was conducted involving 140 randomly selected 10-year-old boys. The mean and standard deviation of their* ***height*** *were found to be 140.8 cm and 7.0 cm, respectively. For* ***body weight****, the mean was 35.6 kg and the standard deviation was also 7.0 kg. Determine which of the two characteristics (height or body weight) shows a greater degree of variation among the boys.*

Solution: Let the mean and standard deviation of height be:

$$\bar{x}_1 = 140.8 \text{ cm}, \quad s_1 = 7.0 \text{ cm}$$

Let the mean and standard deviation of weight be:

$$\bar{x}_2 = 35.6 \text{ kg}, \quad s_2 = 7.0 \text{ kg}$$

For height:

$$CV_{\text{height}} = \left(\frac{7.0}{140.8}\right) \times 100\% = 4.97\% \approx 5.0\%$$

For weight:

$$CV_{\text{weight}} = \left(\frac{7.0}{35.6}\right) \times 100\% = 19.66\% \approx 19.7\%$$

Hence we conclude that the variation in body weight (19.7%) is greater than the variation in height (5.0%) among the 10-year-old boys in this sample.

DRAFT

3.3.7 Exercises

1. Given the following dataset of temperatures (in degrees Celsius) recorded over a week:
$$22, 25, 19, 23, 24, 20, 21$$
   (a) Calculate the range of the temperatures.

2. Consider the following dataset representing the scores of 8 students in a test:
$$75, 85, 95, 80, 90, 70, 88, 92$$
   (a) Compute the variance of the test scores.
   (b) Determine the standard deviation of the test scores.

3. The following dataset represents the frequency distribution of exam scores:
   - Class Intervals:
$$[50-60, 60-70, 70-80, 80-90, 90-100]$$
   - Frequencies:
$$[6, 10, 15, 12, 7]$$
   (a) Calculate the mean of the grouped data.
   (b) Compute the variance and standard deviation of the grouped data.

4. A dataset of the number of hours spent by patients in a hospital (per day) is as follows:
$$1.5, 2.0, 2.5, 3.0, 1.0, 2.2, 2.8$$
   (a) Find the range of the hours spent in the hospital.
   (b) Compute the variance of the dataset.

5. A company tracks the number of units sold each month for the last 6 months:
$$150, 170, 160, 180, 175, 165$$
   (a) Calculate the standard deviation of the units sold.

6. A class of students took two different tests with the following scores:
   - Test 1 Scores:
$$78, 82, 85, 88, 90$$
   - Test 2 Scores:
$$72, 80, 78, 85, 90$$
   (a) Compute the mean and median for both tests.

DRAFT

(b) Compute the variance and standard deviation for both tests.

7. Consider a dataset of monthly income grouped into intervals with corresponding frequencies:

   - Income Intervals:

     $$[2000 - 3000, 3000 - 4000, 4000 - 5000, 5000 - 6000]$$

   - Frequencies:

     $$[5, 8, 12, 6]$$

   (a) Calculate the mean income.

   (b) Compute the variance and standard deviation of the grouped income data.

8. A company records the monthly sales (in thousands of dollars) over the last year:

   $$45, 52, 48, 55, 50, 47, 53, 60, 49, 51, 54, 57$$

   (a) Find the range of the monthly sales data.

   (b) Find the standard deviation of the monthly sales data.

9. Consider the dataset representing the weekly earnings (in dollars) of 5 freelancers:

   $$400, 420, 450, 480, 500$$

   (a) Calculate the mean and standard deviation of the earnings.

   (b) Compute the coefficient of variation (CV) for the earnings dataset.

10. A dataset of exam scores for a class is given as follows:

    $$88, 76, 92, 85, 79, 90, 82$$

    (a) Find the mean and standard deviation of the exam scores.

    (b) Compute the coefficient of variation (CV) for the exam scores.

3.4 Measures of Distribution Shape

Skewness and kurtosis are important descriptive statistics that measure the shape characteristics of a data distribution. Skewness quantifies the asymmetry of the distribution around its mean. Kurtosis, on the other hand, measures the "tailedness" of the distribution. Both measures provide critical insights into the distribution's shape, helping to understand the underlying characteristics of the data.

DRAFT

3.4.1 Skewness

Skewness is a statistical measure that characterizes the **degree of asymmetry of a distribution** around its mean. It indicates whether the data are concentrated more on one side of the mean compared to the other.

Coefficient of Skewness

The coefficient of skewness is a standardized measure of skewness that allows for comparison of the degree of asymmetry between different distributions. It provides insight into the direction and extent of the skew of a data distribution.

There are several coefficients of skewness used to calculate skewness, each providing a different perspective on the skewness of a distribution. Here are two commonly used Coefficients:

(i). Fisher-Pearson Coefficient of Skewness

(ii). Pearson Median Coefficient of Skewness

Fisher-Pearson Coefficient of Skewness

The Fisher-Pearson coefficient measures the asymmetry of the distribution around the mean.

Let us consider a sample $x_1, x_2, \ldots, x_n$ of size n. Then, the sample coefficient of skewness can be defined as follows:

$$\text{Skewness } (Sk) = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{3.1}$$

where,

- n is the number of observations,
- x_i represents each data point,
- $\bar{x}$ is the sample mean,
- s is the sample standard deviation.

The Fisher-Pearson coefficient standardizes the third central moment of the distribution. A skewness of 0 indicates a perfectly symmetrical distribution.

Pearson Median Coefficient of Skewness

The Pearson median coefficient calculates skewness based on the mean and the median. The formula for the Pearson median coefficient of skewness is defined as

$$\text{Skewness } (Sk) = 3\left(\frac{\bar{x} - \text{Median}}{s}\right). \tag{3.2}$$

This method uses the difference between the mean and the median to gauge skewness.

Interpretation

- $Sk = 0$: The distribution is symmetric.
- $Sk > 0$: Positive skewness (right-skewed distribution).
- $Sk < 0$: Negative skewness (left-skewed distribution).

Types of Skewness

Based on the calculated coefficient of skewness, the distribution of data can be categorized as follows:

- **Symmetrical Distribution:** When $Sk = 0$, the data are perfectly symmetrical around the mean. In a symmetric distribution, the mean, median, and mode coincide:

$$\text{Mean} = \text{Median} = \text{Mode}$$

- **Positively Skewed (Right-Skewed):** When $Sk > 0$, the right tail (larger values) of the distribution is longer or fatter than the left tail. The bulk of the data is concentrated on the left. In this case, the mean is usually greater than the median, which is greater than the mode:

$$\text{Mode} < \text{Median} < \text{Mean}$$

- **Negatively Skewed (Left-Skewed):** When $Sk < 0$, the left tail (smaller values) of the distribution is longer or fatter than the right tail. The bulk of the data is concentrated on the right. In this case, the mean is usually less than the median, which is less than the mode:

$$\text{Mean} < \text{Median} < \text{Mode}$$

Problem 3.13. *Consider the following dataset of the number of hours spent studying per day by a group of students 2, 4, 4, 4, 5, 6, 8, and 10 hours. Calculate the skewness.*

DRAFT

Solution

Fisher-Pearson Coefficient of Skewness: The mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+4+4+4+5+6+8+10}{8} = \frac{43}{8} = 5.375 \text{ hours}$$

Using Table 3.6, the variance (s^2) is

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1} x_i^2 - n\bar{x}^2\right) = \frac{1}{8-1}(277 - 8 \times 5.375^2) = \frac{45.875}{7} = 6.5536$$

and hence

$$s = \sqrt{6.5536} = 2.56 \text{ hours}$$

From Table 3.6, we have

$$\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3 = 4.230$$

Table 3.6: Calculations for Skewness

x_i	x_i^2	$\frac{x_i-\bar{x}}{s}$	$\left(\frac{x_i-\bar{x}}{s}\right)^3$
2	4	-1.3184	-2.2914
4	16	-0.5371	-0.1549
4	16	-0.5371	-0.1549
4	16	-0.1465	-0.0031
6	36	0.2441	0.0146
8	64	1.0254	1.0781
10	100	1.8066	5.8968
43	**277**		**4.230**

Therefore,

$$\begin{aligned}
\text{Skewness} &= \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3 \\
&= \frac{8}{(8-1)(8-2)} \times 4.230 \\
&= \frac{8}{42} \times 4.230 \\
&= 0.8057
\end{aligned}$$

The skewness of this dataset is approximately 0.8057. This positive value indicates that the distribution is right-skewed, with a longer tail on the right side.

DRAFT

Pearson Median Coefficient of Skewness: Since the dataset has an even number of observations:

$$\text{Median} = \frac{4+5}{2} = 4.5$$

Hence, the skewness is

$$\text{Skewness } (Sk) = 3\left(\frac{\bar{x} - \text{Median}}{s}\right) = 3\left(\frac{5.375 - 4.5}{2.56}\right) = 1.0254$$

The skewness of the dataset is approximately 1.0254. This positive value indicates that the distribution is right-skewed, meaning it has a longer tail on the right side.

3.4.2 Kurtosis

Even with knowledge of central tendency, dispersion, and skewness, we still do not have a full understanding of a distribution. To gain a complete perspective on the shape of the distribution, we also need to consider **kurtosis**. Kurtosis is a statistical measure that describes the "tailedness" and "peakedness" of the distribution of data. It gives insight into the shape of the distribution compared to the normal distribution.

Karl Pearson's kurtosis is defined as the ratio of the fourth central moment to the square of the variance, also known as the fourth standardized moment of a distribution. The sample version of it is

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}. \tag{3.3}$$

where $x_1, x_2, \ldots, x_n$ be a set of n numerical observations, $\bar{x}$ is the sample mean.

It indicates whether the data are heavy-tailed or light-tailed relative to a normal distribution. The kurtosis of the normal distribution is 3. Therefore, a common tool for measuring kurtosis is the **excess kurtosis**, which is given by

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3.$$

A distribution with **positive excess kurtosis** has a **sharp peak** and **heavy tails**, whereas a distribution with **negative excess kurtosis** has a **flatter peak** and **lighter tails** compared to the normal distribution.

DRAFT

However, the formula give in (3.3) for sample kurtosis is **biased** when estimating the population kurtosis, especially for small samples. To address this, a **bias-corrected sample kurtosis** formula is used, given by

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

where s is the sample standard deviation.

This adjustment helps provide an **unbiased estimate** of the population kurtosis under the assumption of normality, especially for smaller sample sizes. Many software programs (such as Excel's `KURT` function, which uses a bias-corrected formula) use this bias-corrected sample formula to measure the kurtosis.

Let $x_1, x_2, \ldots, x_n$ be a set of n numerical observations. The **bias-corrected sample kurtosis** (Excess Kurtosis) is defined as

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation.

The interpretation of K is given in the Table 3.7.

Table 3.7: Types of kurtosis and their interpretations

Type	Kurtosis	Excess Kurtosis (K)	Description
Mesokurtic	3	0	Normal distribution, moderate tails
Leptokurtic	> 3	Positive	Heavy tails, sharp peak
Platykurtic	< 3	Negative	Light tails, flat peak

Problem 3.14. *Consider Problem 3.13. Calculate the kurtosis of the following dataset, which represents the number of hours spent studying per day by a group of students: 2, 4, 4, 4, 5, 6, 8, and 10 hours.*

Solution

In the solution to Problem 3.13, the mean is $\bar{x} = 5.375$ hours and the standard deviation is $s = 2.56$ hours. The calculation for the kurtosis is provided in Table 3.8.

DRAFT

Table 3.8: Calculations for Kurtosis

x_i	x_i^2	$\frac{x_i-\bar{x}}{s}$	$\left(\frac{x_i-\bar{x}}{s}\right)^4$
2	4	-1.3184	3.0209
4	16	-0.5371	0.0832
4	16	-0.5371	0.0832
4	16	-0.5371	0.0832
5	25	-0.1465	0.0005
6	36	0.2441	0.0036
8	64	1.0254	1.1055
10	100	1.8066	10.6535
43	**277**		**15.0336**

Using above table, we have

$$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 = 15.03358$$

Therefore,

$$\begin{aligned} K &= \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \\ &= \frac{8\times 9}{7\times 6\times 5}\times 15.03358 - \frac{3\times 7^2}{6\times 5} \\ &= 0.343\times 15.03358 - 4.9 \\ &= 0.2543 \end{aligned}$$

The excess kurtosis of the dataset is approximately 0.2543, the positive value suggests that the distribution is leptokurtic. This means the distribution has heavier tails and is more peaked compared to a normal distribution, indicating a higher probability of extreme values or outliers.

3.4.3 Exercises

1. Given the following dataset representing the monthly number of new customer sign-ups for a company:

   $$12, 15, 14, 16, 21, 25, 30, 35, 40, 50$$

   Calculate the skewness of the dataset. Use a statistical software or formula for skewness calculation.

DRAFT

2. The following dataset represents the heights (in cm) of 10 students:

$$150, 155, 160, 165, 170, 175, 180, 185, 190, 195$$

Determine the kurtosis of the height dataset. Use a statistical software or formula for kurtosis calculation.

3. A company tracks monthly sales figures (in thousands of dollars) for a year:

$$40, 42, 44, 45, 47, 50, 52, 55, 60, 65, 70, 75$$

(a) Calculate the skewness of the sales data.

(b) Determine the kurtosis of the sales data.

4. The daily maximum temperatures (in degrees Celsius) for a week are:

$$18, 20, 22, 24, 26, 28, 30$$

(a) Compute the skewness of the temperature data.

(b) Determine the interquartile range (IQR) of the temperature data.

3.5 Measures of Position

While measures of central tendency (like the mean or median) describe the 'average' and measures of dispersion (like the standard deviation) describe the 'spread', **measures of position** are designed to describe the **relative standing** of a specific data point within a dataset. In essence, they answer the question: "Where does this particular value fall compared to the rest of the data?"

For data scientists, these measures are indispensable tools for **Exploratory Data Analysis (EDA)**, identifying **outliers**, feature scaling, and understanding data distribution. The primary measures of position we will focus on are **Quartiles**, **Percentiles**, and **Z-Scores**.

Quartiles and percentiles provide essential tools for summarizing data by dividing it into meaningful segments. These measures enable data scientists to capture the underlying structure and variability, highlighting not only central tendencies but also the dispersion and skewness within datasets.

A critical application of these measures is **outlier detection**. Outliers are data points that deviate markedly from the rest of the observations, often indicating valuable insights or potential data quality issues. Identifying these points prevents misleading conclusions and guides appropriate data preprocessing steps.

DRAFT

The statistical summaries derived from these measures are often synthesized in a powerful graphical representation: the **Boxplot**. Boxplots integrate quartiles, percentiles, and outlier information into a compact visual summary, allowing analysts to quickly grasp key aspects of the data distribution, such as central tendency, variability, and asymmetry.

In this section, we will first define and explore **Quartiles and Percentiles**, examining their direct role in detecting outliers via the **Interquartile Range (IQR)**. We will then cover **Z-Scores** as a standardization technique for measuring relative standing across different distributions. Finally, we will show how **Boxplots** provide a clear and efficient means of visualizing these statistical summaries, making them indispensable tools in the data science toolkit.

3.5.1 Quartiles

In data exploration, quartiles are essential tools to summarize and describe the distribution of a dataset. It provides a concise summary of a data set, highlighting its central tendency and variability without requiring a full description of the data. They help identify the spread of the data, allowing you to see how values are distributed. Quartiles divide a data set into four equal parts, each representing 25% of the data. The range between the first quartile (Q_1) and the third quartile (Q_3) is known as the interquartile range (IQR), which indicates where the middle 50% of the data lies.

Suppose we have a dataset of $x_1, x_2, \ldots, x_n$ of size n, and let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered version of this dataset, where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

To locate the position of the kth quartile, we define:

$$m_k = \frac{n+1}{4} \times k,$$

where $k = 1, 2, 3$ correspond to the first, second, and third quartiles, respectively. Since the position m_k may not always be an integer, we use linear interpolation to determine the value of the quartile.

The quartile Q_k is calculated as follows:

$$Q_k = \begin{cases} x_{(m_k)}, & \text{if } m_k \text{ is an integer,} \\ x_{(\lfloor m_k \rfloor)} + f_k \cdot \left(x_{(\lceil m_k \rceil)} - x_{(\lfloor m_k \rfloor)} \right), & \text{otherwise,} \end{cases} \tag{3.4}$$

where,

- $\lfloor m_k \rfloor$ is the greatest integer less than or equal to m_k.

DRAFT

- $\lceil m_k \rceil$ is the smallest integer greater than or equal to m_k.
- $f_k = m_k - \lfloor m_k \rfloor$ represents the fractional part of m_k.

If m_k is an integer, no interpolation is necessary, and the quartile value Q_k corresponds directly to the data point at that position in the ordered dataset. However, if m_k is not an integer, Q_k is computed by linearly interpolating between the two closest ordered data values.

This method is an example of **linear interpolation** between the data values to estimate the quartiles, providing more precise quantile values when the dataset size does not match exactly with the desired integer index positions.

Remark 3.5.1. ***Ceiling Function:*** *The ceiling function, denoted as* $\lceil m_k \rceil$*, is defined as:*

$$\lceil m_k \rceil = \textit{the smallest integer greater than or equal to } m_k$$

Floor Function: *The floor function, denoted as* $\lfloor m_k \rfloor$*, is defined as:*

$$\lfloor m_k \rfloor = \textit{the largest integer less than or equal to } m_k$$

For example, for the floor function: $\lfloor 2.4 \rfloor = 2$*, and for the ceiling function:* $\lceil 2.4 \rceil = 3$.

Problem 3.15. *Suppose that a college placement office sent a questionnaire to a sample of business school graduates requesting information on monthly starting salaries. Table 3.9 shows the collected data.*

Table 3.9: Monthly Starting Salaries for a Sample of 12 Business School Graduates (4-column Format)

Graduate	Salary ($)	Graduate	Salary ($)
1	5850	7	5890
2	5950	8	6130
3	6050	9	5940
4	5880	10	6325
5	5755	11	5920
6	5710	12	5880

Compute the quartiles of the monthly starting salary for the sample of 12 business college graduates.

DRAFT

Solution

To compute the quartiles, we first sort the salaries in ascending order:

$$\begin{aligned} &x_{(1)} = 5710, \quad x_{(2)} = 5755, \quad x_{(3)} = 5850, \quad x_{(4)} = 5880, \\ &x_{(5)} = 5880, \quad x_{(6)} = 5890, \quad x_{(7)} = 5920, \quad x_{(8)} = 5940, \\ &x_{(9)} = 5950, \quad x_{(10)} = 6050, \quad x_{(11)} = 6130, \quad x_{(12)} = 6325. \end{aligned}$$

Quartile Computations

- **First Quartile (Q1):**

$$m_1 = \frac{13 \times 1}{4} = 3.25$$

where $\lfloor 3.25 \rfloor = 3$, $\lceil 3.25 \rceil = 4$, $f_1 = 3.25 - 3 = 0.25$. Substituting values, we have

$$\begin{aligned} Q_1 &= x_{(3)} + 0.25 \cdot \left(x_{(4)} - x_{(3)}\right) \\ &= 5850 + 0.25 \times (5880 - 5850) \\ &= 5857.5 \end{aligned}$$

- **Second Quartile (Q2):**

$$m_2 = \frac{13 \times 2}{4} = 6.5$$

where $\lfloor 6.5 \rfloor = 6$, $\lceil 6.5 \rceil = 7$, $f_2 = 6.5 - 6 = 0.5$. Substituting values, we have

$$\begin{aligned} Q_2 &= x_{(6)} + 0.5 \times (x_{(7)} - x_{(6)}) \\ &= 5890 + 0.5 \times (5920 - 5890) \\ &= 5905 \end{aligned}$$

- **Third Quartile (Q3):**

$$m_3 = \frac{13 \times 3}{4} = 9.75$$

where $\lfloor 9.75 \rfloor = 9$, $\lceil 9.75 \rceil = 10$, $f_3 = 9.75 - 9 = 0.75$. Substituting values, we have

$$\begin{aligned} Q_3 &= x_{(9)} + 0.75 \cdot \left(x_{(10)} - x_{(9)}\right) \\ &= 5950 + 0.75 \times (6050 - 5950) \\ &= 6025 \end{aligned}$$

The quartiles for the monthly starting salaries are:

DRAFT

- 1st Quartile (Q_1): 5857.5
- Second Quartile (Q2) or Median: 5905
- 3rd Quartile (Q_3): 6025

3.5.2 Deciles

Deciles divide a dataset into ten equal parts. The dth decile D_d represents the value below which $d \times 10\%$ of the observations lie. Like quartiles, deciles can be estimated using linear interpolation on the ordered dataset.

Let $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ denote the ordered dataset, where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

To determine the position of the dth decile, we calculate:

$$m_d = \frac{n+1}{10} \times d,$$

where $1 \leq d \leq 9$ specifies the desired decile.

If m_d is an integer, the decile D_d is given directly by:

$$D_d = x_{(m_d)}.$$

Otherwise, linear interpolation is used:

$$D_d = x_{(\lfloor m_d \rfloor)} + f_d \cdot \left(x_{(\lceil m_d \rceil)} - x_{(\lfloor m_d \rfloor)}\right),$$

where:

$$f_d = m_d - \lfloor m_d \rfloor.$$

If m_d is not an integer, the decile value D_d lies between two adjacent ordered data points. The fractional part f_d represents the contribution of the next data point in the interpolation, ensuring a smooth and accurate estimate of the desired decile.

Problem 3.16. *Refer to the Problem 3.15, the ordered annual salaries (in dollars) of 12 employees:*

$$5710, \quad 5755, \quad 5850, \quad 5880, \quad 5880, \quad 5890,$$
$$5920, \quad 5940, \quad 5950, \quad 6050, \quad 6130, \quad 6325,$$

Calculate the 1st decile (D_1), 5th decile (D_5), and 9th decile (D_9) using interpolation.

DRAFT

solution: The dataset size is $n = 12$. The position m_d for the dth decile is:

$$m_d = \frac{n+1}{10} \times d = \frac{13}{10} \times d = 1.3\,d.$$

- **1st decile ($d = 1$):**
  We find that $m_1 = 1.3 \times 1 = 1.3$, so $\lfloor m_1 \rfloor = 1$, $\lceil m_1 \rceil = 2$, and $f_1 = 0.3$. Thus,

$$\begin{aligned} D_1 &= x_{(1)} + 0.3(x_{(2)} - x_{(1)}) \\ &= 5710 + 0.3(5755 - 5710) \\ &= 5723.5. \end{aligned}$$

- **5th decile ($d = 5$):**
  We calculate $m_5 = 1.3 \times 5 = 6.5$, so $\lfloor m_5 \rfloor = 6$, $\lceil m_5 \rceil = 7$, and $f_5 = 0.5$. Therefore,

$$\begin{aligned} D_5 &= x_{(6)} + 0.5(x_{(7)} - x_{(6)}) \\ &= 5890 + 0.5(5920 - 5890) \\ &= 5905. \end{aligned}$$

- **9th decile ($d = 9$):**
  We calculate $m_9 = 1.3 \times 9 = 11.7$, so $\lfloor m_9 \rfloor = 11$, $\lceil m_9 \rceil = 12$, and $f_9 = 0.7$. Thus,

$$\begin{aligned} D_9 &= x_{(11)} + 0.7(x_{(12)} - x_{(11)}) \\ &= 6130 + 0.7(6325 - 6130) \\ &= 6266.5. \end{aligned}$$

Therefore, the 1st, 5th, and 9th deciles of the salaries are approximately \$5723.5, \$5905, and \$6266.5, respectively.

3.5.3 Percentiles

Percentiles are a generalization of quartiles that divide a dataset into 100 equal parts. The pth percentile P_p represents the value below which $p\%$ of the observations lie. This can be determined using a similar interpolation method based on the ordered dataset.

Let $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ denote the ordered dataset, where

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

To locate the position of the pth percentile, we define:

$$m_p = \frac{n+1}{100} \times p,$$

DRAFT

where $1 \leq p \leq 100$ represents the desired percentile. If m_p is an integer, the percentile P_p is directly given by:

$$P_p = x_{(m_p)}.$$

Otherwise, linear interpolation is used:

$$P_p = x_{(\lfloor m_p \rfloor)} + f_p \cdot \left(x_{(\lceil m_p \rceil)} - x_{(\lfloor m_p \rfloor)} \right),$$

where:

$$f_p = m_p - \lfloor m_p \rfloor.$$

If m_p is not an integer, P_p lies between the two adjacent ordered values. The fractional part f_p represents the weight of the difference between these two values in determining the percentile. This linear interpolation approach provides a precise and consistent estimate of the desired percentile, ensuring accurate characterization of the dataset's distribution.

Problem 3.17. *Refer to the Problem 3.15, a company has recorded the annual salaries (in dollars) of 12 employees:*

$$5710, \quad 5755, \quad 5850, \quad 5880, \quad 5880, \quad 5890,$$

$$5920, \quad 5940, \quad 5950, \quad 6050, \quad 6130, \quad 6325.$$

Determine the 10th percentile, 50th percentile (median), and 90th percentile using the linear interpolation formula for percentiles.

Solution: First, we order the salaries in ascending order:

$$\begin{aligned} &x_{(1)} = 5710, \quad x_{(2)} = 5755, \quad x_{(3)} = 5850, \quad x_{(4)} = 5880, \\ &x_{(5)} = 5880, \quad x_{(6)} = 5890, \quad x_{(7)} = 5920, \quad x_{(8)} = 5940, \\ &x_{(9)} = 5950, \quad x_{(10)} = 6050, \quad x_{(11)} = 6130, \quad x_{(12)} = 6325. \end{aligned}$$

The dataset size is $n = 12$. The position m_p for the pth percentile is calculated as

$$m_p = \frac{n+1}{100} \times p = \frac{13}{100} \times p = 0.13\, p.$$

- **10th percentile ($p = 10$):**
  We calculate $m_{10} = 0.13 \times 10 = 1.3$, so $\lfloor m_{10} \rfloor = 1$, $\lceil m_{10} \rceil = 2$, and $f_{10} = 0.3$. Using these values:

$$\begin{aligned} P_{10} &= x_{(1)} + 0.3(x_{(2)} - x_{(1)}) \\ &= 5710 + 0.3(5755 - 5710) \\ &= 5710 + 0.3 \times 45 \\ &= 5710 + 13.5 = 5723.5. \end{aligned}$$

DRAFT

- **50th percentile ($p = 50$):**
  We calculate $m_{50} = 0.13 \times 50 = 6.5$, so $\lfloor m_{50} \rfloor = 6$, $\lceil m_{50} \rceil = 7$, and $f_{50} = 0.5$. Thus,

$$\begin{aligned} P_{50} &= x_{(6)} + 0.5(x_{(7)} - x_{(6)}) \\ &= 5890 + 0.5(5920 - 5890) \\ &= 5905. \end{aligned}$$

- **90th percentile ($p = 90$):**
  We calculate $m_{90} = 0.13 \times 90 = 11.7$, so $\lfloor m_{90} \rfloor = 11$, $\lceil m_{90} \rceil = 12$, and $f_{90} = 0.7$. Thus,

$$\begin{aligned} P_{90} &= x_{(11)} + 0.7(x_{(12)} - x_{(11)}) \\ &= 6130 + 0.7(6325 - 6130) \\ &= 6266.5. \end{aligned}$$

Therefore, the 10th, 50th, and 90th percentiles of the sample salaries are approximately \$5723.5, \$5905, and \$6266.5, respectively.

3.5.4 Interquartile Range (IQR)

The IQR measures the range within which the middle 50% of the data falls, from the first quartile (Q_1) to the third quartile (Q_3). Hence, the interquartile range is defined as

$$\text{IQR} = Q_3 - Q_1$$

Consider a dataset of test scores:

$$55, 60, 65, 70, 75, 80, 85, 90, 95, 100.$$

The dataset size is $n = 10$. The position of the first quartile is

$$m_1 = \frac{n+1}{4} = \frac{11}{4} = 2.75.$$

Using the equation given (3.4)

$$Q_1 = x_{(2)} + 0.75 \times \left(x_{(3)} - x_{(2)}\right) = 60 + 0.75 \times (65 - 60) = 63.75.$$

The position of the third quartile is

$$m_3 = \frac{3(n+1)}{4} = \frac{33}{4} = 8.25.$$

Using interpolation,

$$Q_3 = x_{(8)} + 0.25 \times \left(x_{(9)} - x_{(8)}\right) = 90 + 0.25 \times (95 - 90) = 91.25.$$

The interquartile range (IQR) is

$$\text{IQR} = Q_3 - Q_1 = 91.25 - 63.75 = 27.5.$$

DRAFT

3.5.5 Outlier Identification via IQR Rule

In statistics, an **outlier** is a data point that is significantly different from other observations. We can detect outliers using the interquartile range (IQR). Given a dataset $x_1, x_2, \ldots, x_n$, the outlier detection can be described as follows:

Define Mild and Extreme Outliers:

- Mild Outlier: A data point x_i is considered a mild outlier if

$$x_i < Q_1 - 1.5 \times \text{IQR} \quad \text{or} \quad x_i > Q_3 + 1.5 \times \text{IQR}$$

- Extreme Outlier: A data point x_i is considered an extreme outlier if

$$x_i < Q_1 - 3 \times \text{IQR} \quad \text{or} \quad x_i > Q_3 + 3 \times \text{IQR}$$

If there are no mild outliers in a dataset (i.e., no data points fall outside the fences defined by $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$), then there cannot be any extreme outliers either. This is because the extreme outlier fences ($Q_1 - 3 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$) lie even further out from the data's center. Therefore, the absence of mild outliers ensures the absence of extreme outliers.

Problem 3.18. *Consider the following dataset representing the heights (in cm) of a sample of adult men:*

$$172,\ 167,\ 168,\ 169,\ 170,\ 171,\ 173,\ 175,\ 177,\ 200, 175$$

Calculate the first quartile (Q_1), second quartile (Q_2), and third quartile (Q_3) using the interpolation formulas. Then determine any mild and extreme outliers.

Solution: First, we order the dataset in ascending order:

$$x_{(1)} = 165,\ x_{(2)} = 167,\ x_{(3)} = 168,\ x_{(4)} = 169,\ x_{(5)} = 170,\ x_{(6)} = 171,$$

$$x_{(7)} = 172,\ x_{(8)} = 173,\ x_{(9)} = 175,\ x_{(10)} = 177,\ x_{(11)} = 200$$

The sample size is $n = 11$.

- **First Quartile (Q_1):**

$$m_1 = \frac{n+1}{4} \times 1 = \frac{12}{4} = 3$$

Since $m_1 = 3$ is an integer:

$$Q_1 = x_{(3)} = 168$$

DRAFT

- **Second Quartile (Q_2):**

$$m_2 = \frac{n+1}{4} \times 2 = \frac{12}{4} \times 2 = 6$$

Since $m_2 = 6$ is an integer:

$$Q_2 = x_{(6)} = 171$$

- **Third Quartile (Q_3):**

$$m_3 = \frac{n+1}{4} \times 3 = \frac{12}{4} \times 3 = 9$$

Since $m_3 = 9$ is an integer:

$$Q_3 = x_{(9)} = 175$$

Interquartile Range (IQR):

$$\text{IQR} = Q_3 - Q_1 = 175 - 168 = 7$$

Mild Outlier Thresholds:

$$\text{Lower mild threshold} = Q_1 - 1.5 \times \text{IQR} = 168 - 1.5 \times 7 = 168 - 10.5 = 157.5$$
$$\text{Upper mild threshold} = Q_3 + 1.5 \times \text{IQR} = 175 + 10.5 = 185.5$$

Extreme Outlier Thresholds:

$$\text{Lower extreme threshold} = Q_1 - 3 \times \text{IQR} = 168 - 21 = 147$$
$$\text{Upper extreme threshold} = Q_3 + 3 \times \text{IQR} = 175 + 21 = 196$$

Observing the data, any values above the upper mild threshold of 185.5 are considered mild outliers, while values exceeding the upper extreme threshold of 196 are classified as extreme outliers. Since 200 is greater than both 185.5 and 196, it is identified as both a mild and extreme outlier.

□

Problem 3.19. *Consider the following systolic blood pressure (SBP) readings (in mmHg):*

$$165, 50, 110, 120, 125, 130, 135, 140, 145, 150, 155,$$
$$160, 175, 180, 185, 190, 195, 200, 115, 220, 170$$

Calculate the first and third quartiles and then determine mild and extreme outliers.

Solution: First, we order the data in ascending order:

$$50, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155,$$
$$160, 165, 170, 175, 180, 185, 190, 195, 200, 220$$

The number of data points is $n = 21$.

DRAFT

Finding Q_1:

$$m_1 = \frac{n+1}{4} \times 1 = \frac{21+1}{4} = 5.5,$$

and

$$Q_1 = x_{(\lfloor 5.5 \rfloor)} + f_1 \cdot (x_{(\lceil 5.5 \rceil)} - x_{(\lfloor 5.5 \rfloor)})$$

where, $\lfloor 5.5 \rfloor = 5$, $\lceil 5.5 \rceil = 6$, $f_1 = 5.5 - 5 = 0.5$. Thus, the first quartile is

$$Q_1 = 125 + 0.5 \cdot (130 - 125) = 125 + 0.5 \cdot 5 = 125 + 2.5 = 127.5$$

Finding Q_3:

For the third quartile, we calculate

$$m_3 = \frac{n+1}{4} \times 3 = \frac{21+1}{4} \times 3 = \frac{22}{4} \times 3 = 16.5$$

and

$$Q_3 = x_{(\lfloor 16.5 \rfloor)} + f_3 \cdot (x_{(\lceil 16.5 \rceil)} - x_{(\lfloor 16.5 \rfloor)})$$

where $\lfloor 16.5 \rfloor = 16$, $\lceil 16.5 \rceil = 17$, $f_3 = 16.5 - 16 = 0.5$. Hence,

$$Q_3 = 180 + 0.5 \cdot (185 - 180) = 180 + 0.5 \cdot 5 = 180 + 2.5 = 182.5$$

So, the interquartile range is

$$\text{IQR} = Q_3 - Q_1 = 182.5 - 127.5 = 55$$

Determine Outlier Thresholds:

- Mild Outliers:

$$\begin{aligned} \text{Lower Mild Threshold} &= Q_1 - 1.5 \times \text{IQR} = 127.5 - 1.5 \times 55 \\ &= 127.5 - 82.5 = 45 \end{aligned}$$

$$\begin{aligned} \text{Upper Mild Threshold} &= Q_3 + 1.5 \times \text{IQR} = 182.5 + 1.5 \times 55 \\ &= 182.5 + 82.5 = 265 \end{aligned}$$

Mild outliers are values below 45 or above 265. Since all data points lie between 50 and 220, which is within these mild outlier Thresholds, there are no mild outliers in SBP readings. We know if there are no mild outliers, there cannot be any extreme outliers either. Therefore, there are no outliers in this dataset. □

DRAFT

3.5.6 Python Code: Dispersion Measures

In this section, we will demonstrate how to compute the range, variance, standard deviation, coefficient of variation, quartiles, interquartile range (IQR), skewness, and kurtosis using Python. Consider the following dataset:

$$\{10, 15, 15, 20, 25, 30\}$$

```python
import numpy as np
from scipy.stats import iqr, skew, kurtosis

# Sample data
data = [10, 15, 15, 20, 25, 30]

# Range
data_range = np.ptp(data)
print(f"Range: {data_range}")

# Variance
variance = np.var(data, ddof=1)
print(f"Variance: {variance}")

# Standard Deviation
std_deviation = np.std(data, ddof=1)
print(f"Standard Deviation: {std_deviation}")

# Coefficient of Variation
coef_of_variation = std_deviation / np.mean(data)
print(f"Coefficient of Variation: {coef_of_variation}")

# Quartiles
quartiles = np.percentile(data, [25, 50, 75])
print(f"Quartiles (25th, 50th, 75th): {quartiles}")

# Interquartile Range (IQR)
interquartile_range = iqr(data)
print(f"Interquartile Range (IQR): {interquartile_range}")

# Skewness
data_skewness = skew(data)
print(f"Skewness: {data_skewness}")

# Kurtosis
data_kurtosis = kurtosis(data)
print(f"Kurtosis: {data_kurtosis}")
```

The output of the code will be:

- **Range:** 20

- **Variance:** 62.5
- **Standard Deviation:** 7.91
- **Coefficient of Variation:** 0.42
- **Quartiles (25th, 50th, 75th):** [15. 17.5 25.]
- **Interquartile Range (IQR):** 10.0
- **Skewness:** 0.246
- **Kurtosis:** -1.09

3.5.7 Z-Scores (Standard Scores)

The **Z-score**, or **Standard Score**, measures how many **standard deviations** a specific data point is away from the **mean** of the dataset. This measure is used for **standardization**, which allows us to compare values from different datasets that may have different units, means, and standard deviations.

The formula for the Z-score (z) of a data point (x) is:

$$z = \frac{x - \mu}{\sigma}$$

Where,

- x is the individual data point.
- μ is the population mean.
- σ is the population standard deviation.

Applications of Z-score

- **Outlier Detection:** If the data are normally distributed, any value with a Z-score greater than 3 or less than -3 is usually considered an outlier, as these points fall outside of approximately 99.7% of the data (based on the Empirical Rule).
- **Data Standardization:** Z-scores are the result of the `StandardScaler` transformation in machine learning. Standardization places all features on a consistent scale (mean of 0 and standard deviation of 1), which is essential for distance-based algorithms like k-Nearest Neighbors (kNN) and for models using gradient descent.

DRAFT

3.5.8 Five-Number Summary

The five-number summary is a concise descriptive statistic that provides important information about the distribution of a dataset. It consists of the following five values:

- **Minimum** ($x_{(1)}$): The smallest value in the dataset.
- **First Quartile** (Q_1): The value below which 25% of the data fall.
- **Median** (Q_2): The middle value that divides the dataset into two equal halves.
- **Third Quartile** (Q_3): The value below which 75% of the data fall.
- **Maximum** ($x_{(n)}$): The largest value in the dataset.

The five-number summary provides a quick overview of the center, spread, and range of the data, and is often used to create boxplots for visualizing distributions.

Example: Refer to the Problem 3.15, we found

- Minimum: 5710
- First Quartile: $Q_1 = 5857.5$
- Median: $Q_2 = 5905$
- Third Quartile: $Q_3 = 6025$
- Maximum: 6325

Thus, the five-number summary is

$$(5710, \quad 5857.5, \quad 5905, \quad 6025, \quad 6325).$$

3.5.9 Boxplot

In the field of data science, understanding the distribution and variability of data is essential for meaningful analysis and decision-making. Boxplots, also known as box-and-whisker plots, are powerful graphical tools that provide a concise yet comprehensive summary of a dataset's key characteristics. They are graphical representations of the five-number summary of a dataset.

By offering a visual snapshot of central tendency, spread, skewness, and outliers, boxplots enable data scientists to quickly assess the underlying patterns and potential anomalies in their data.

DRAFT

Boxplot: A boxplot graphically represents the five-number summary of a dataset: minimum, first quartile, median, third quartile, and maximum. It visually summarizes the dataset's central tendency, dispersion, and skewness, and it also highlights the presence of outliers.

The key components of a boxplot are

- **Box:** The box spans from the first quartile (Q_1) to the third quartile (Q_3) and represents the interquartile range (IQR), which contains the middle 50% of the data.
- **Median line:** A line inside the box indicates the median (Q_2) of the data, showing the dataset's central tendency.
- **Whiskers:** Lines extending from the box to the smallest and largest data points within 1.5 times the IQR from the quartiles. They represent the range of typical data values. The choice of the factor 1.5 is a practical balance suggested by John Tukey. This rule remains widely adopted due to its empirical effectiveness across a variety of data types and disciplines.
- **Outliers:** Points plotted beyond the whiskers are considered outliers. Mild outliers lie between 1.5 and 3 times the IQR from the quartiles, while extreme outliers lie beyond 3 times the IQR.

The Figure 3.1 below illustrates these components:

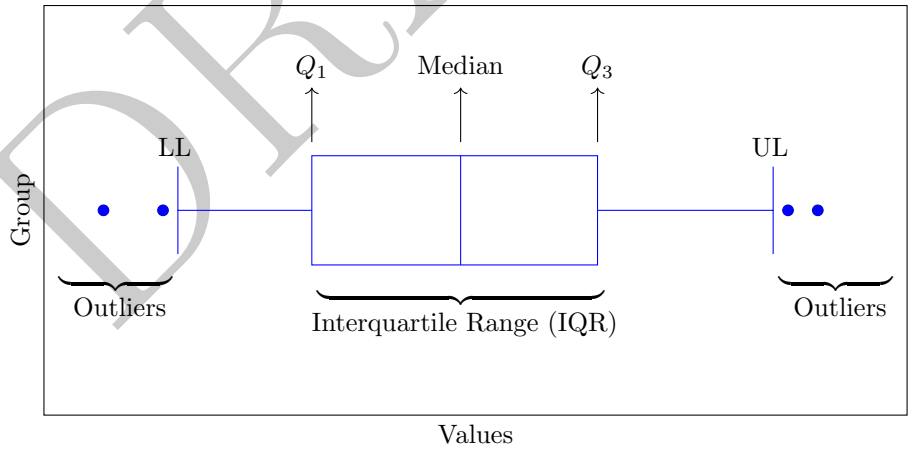


Figure 3.1: Detailed boxplot illustrating the distribution of a sample dataset with components: LL (Lower whisker), Q1, Median, Q3, UL (Upper whisker), and Outliers, where LL $= Q_1 - 1.5 \times$ IQR and UL $= Q_3 + 1.5 \times$ IQR

DRAFT

The concept of the boxplot was introduced by the renowned statistician John Tukey in the 1970s as a simple and effective way to visualize data distribution. One of Tukey's key contributions is the rule of using **1.5 times the interquartile range (IQR)**, where IQR $= Q3 - Q1$, to identify potential outliers. This multiplier serves as a robust boundary to distinguish between typical data variation and unusually extreme values. Data points lying beyond the fences defined as $Q1 - 1.5 \times \text{IQR}$ and $Q3 + 1.5 \times \text{IQR}$ are flagged as outliers, highlighting observations that warrant further investigation.

Why is the boxplot so important?

Boxplots are an invaluable tool in data science because they provide a succinct and insightful visual summary of key characteristics of a dataset. They allow for comparisons between different groups or datasets. Specifically, they help data scientists quickly understand:

- **Central Tendency:** The median (the line inside the box) provides a clear indication of the central value of the data.
- **Spread/Dispersion:** The length of the box (Interquartile Range – IQR) shows how spread out the middle 50% of the data is. The whiskers extend to indicate the overall range of the data within a certain limit (usually 1.5 times the IQR), giving a sense of the spread beyond the central portion.
- **Skewness:** The position of the median within the box and the relative lengths of the whiskers on either side can suggest whether the data is symmetric, positively skewed (longer whisker/tail to the right), or negatively skewed (longer whisker/tail to the left).
- **Outliers:** Individual points plotted beyond the whiskers are identified as potential outliers, which are data points significantly different from the rest of the dataset. These often warrant further investigation as they could be errors or genuine anomalies.

Interpretation tips: When interpreting a boxplot, several key aspects should be considered to understand the distribution and variability of the data.

- **First**, examine the central box (the IQR), which shows where the middle 50% of data falls, and the line inside the box is the median.
- **Secondly**, assess the skewness by examining the median's position within the box: if it is centered, the data is symmetric; if it is closer to one end, the data is skewed.
- **Thirdly**, identify outliers (dots outside the whiskers), which highlight potential variability or anomalies in the data; the number of outliers can also indicate data dispersion or irregularities.

DRAFT

- **Fourth**, observe the box width and whisker lengths to assess the data's spread and dispersion.
- **Finally**, comparing boxplots side-by-side helps identify differences in median, IQR, and outliers, providing insights into data patterns and trends.

Remark 3.5.2. *The lines extending from either end of the box are called* ***whiskers****. The term* ***whisker plot*** *is often used interchangeably with "****box-plot****", focusing on the whiskers. It highlights the range of data within 1.5 times the IQR from the quartiles but might not always show the box or median. Both terms generally refer to the same plot, but "boxplot" is the more comprehensive term, including the full depiction of the quartiles and median along with the whiskers.*

Problem 3.20. *Refer to Problem 3.15. The monthly starting salaries, sorted in ascending order, are:*

$$5710, 5755, 5850, 5880, 5880, 5890, 5920, 5940, 5950, 6050, 6130, 6325$$

The five-number summary for the monthly starting salaries is as follows:

- *Minimum: 5710*
- *1st Quartile (Q_1): 5857.5*
- *Median (Q_2): 5905*
- *3rd Quartile (Q_3): 6025*
- *Maximum: 6325*

Draw the boxplot to represent this data and comment on your findings.

Solution

From the solution of 3.15, the five-number summary for the monthly starting salaries is as follows:

- Minimum: 5710
- 1st Quartile (Q_1): 5857.5
- Median (Q_2): 5905
- 3rd Quartile (Q_3): 6025
- Maximum: 6325

DRAFT

To find the lower and upper whiskers (bounds), we first calculate the interquartile range (IQR):

$$\text{IQR} = Q_3 - Q_1 = 6025 - 5857.5 = 167.5$$

Using 1.5 times the IQR, we can calculate the lower and upper whiskers as follows:

$$\text{Lower whisker} = Q_1 - 1.5 \times \text{IQR} = 5857.5 - 1.5 \times 167.5 = 5606.25$$

$$\text{Upper whisker} = Q_3 + 1.5 \times \text{IQR} = 6025 + 1.5 \times 167.5 = 6276.25$$

Checking for outliers

- Lower outlier: Any value below 5606.25. There are no values below 5710, so there are no lower outliers.
- Upper outlier: Any value above 6276.25. The value 6325 is greater than 6276.25, so 6325 is an upper outlier.

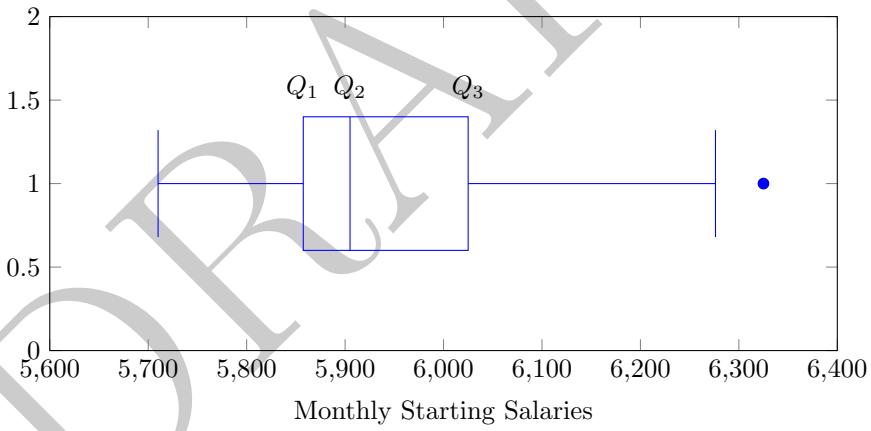


Figure 3.2: Boxplot of monthly starting salaries

Comment on the Boxplot: The boxplot of the monthly starting salaries reveals a **positively skewed** distribution. The median ($Q_2 = 5905$) is positioned slightly closer to the lower quartile, and the right whisker is longer than the left whisker, which indicates that there are some higher salaries pulling the data to the right.

The **lower whisker** extends to **5710**, while the **upper whisker** reaches **6130**. The interquartile range (IQR), between $Q_1 = 5857.5$ and $Q_3 = 6025$, is fairly narrow, indicating that the middle 50% of the data are clustered together.

DRAFT

However, the longer upper whisker and the presence of an outlier at **6325** show that there are some higher salaries that deviate from the general pattern.

The larger portion of the box on the right side of the median line, along with the outlier (**6325**), clearly indicates a **positive skewness** in the data. This suggests that while most starting salaries fall within a consistent range, a few are considerably higher. □

Problem 3.21. *We consider the following data on the cholesterol levels (in mg/dL) of 15 patients:*

$$180, 195, 170, 200, 210, 175, 205, 190, 195, 220, 185, 215, 200, 190, 275$$

Calculate the five-number summary and draw a boxplot to represent this data.

Solution

Given the ordered cholesterol levels:

$$170, 175, 180, 185, 190, 190, 195, 195, 200, 200, 205, 210, 215, 220, 275$$

- Minimum = 170
- Maximum = 275
- Median (Q_2): Since there are 15 data points (odd), the median is the middle value:

$$Q_2 = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{15+1}{2}\right)} = x_{(8)} = 195$$

- First Quartile (Q_1):

$$m = \frac{n+1}{4} = \frac{15+1}{2} = 4$$

The 4th value is:

$$Q_1 = x_{(4)} = 185$$

- Third Quartile (Q_3):

$$m = \frac{15+1}{4} \times 3 = 12$$

Therefore,

$$Q_3 = x_{(12)} = 210$$

- Calculate the IQR:

$$\text{IQR} = Q_3 - Q_1 = 210 - 185 = 25$$

DRAFT

- Calculate the Lower whisker:

  $$\text{Lower whisker} = Q_1 - 1.5 \times \text{IQR} = 185 - 1.5 \times 25 = 185 - 37.5 = 147.5$$

- Calculate the Upper whisker:

  $$\text{Upper whisker} = Q_3 + 1.5 \times \text{IQR} = 210 + 1.5 \times 25 = 210 + 37.5 = 247.5$$

  The value 275 is considered an outlier since it is greater than the upper whisker of 247.5.

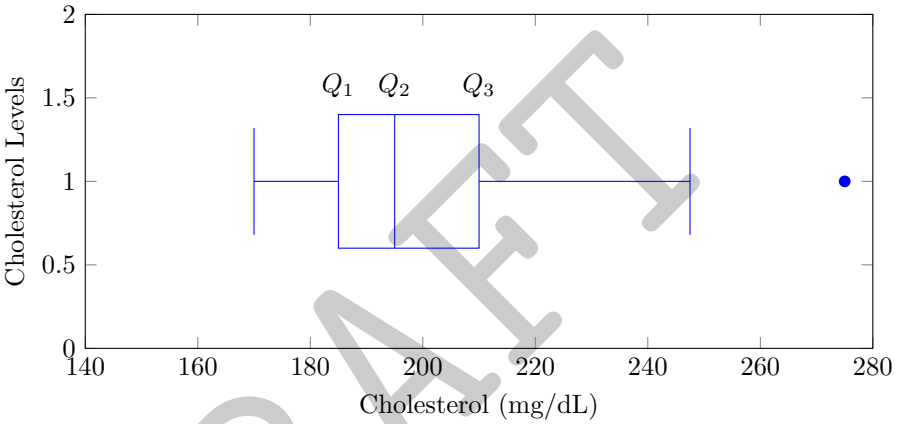


Figure 3.3: Boxplot of Cholesterol Levels

DRAFT

The boxplot of cholesterol levels given in Figure 3.3, provides a clear visual representation of the data distribution. The spread of the data is captured by the range from the minimum value (170 mg/dL) to the maximum value (225 mg/dL), as well as by the interquartile range (IQR) of 25 mg/dL, which measures the spread of the middle 50% of the data between the first quartile (185 mg/dL) and the third quartile (210 mg/dL). There is an outlier in this dataset which is 275. The median (195 mg/dL) lies closer to the first quartile, suggesting a slight right skewness in the data, as the upper quartile range is wider than the lower quartile range. Overall, the data appears relatively symmetric but with a mild skew to the right, indicating that higher cholesterol values are slightly more spread out than the lower values.

3.5.10 Python Code: Boxplot

To create a boxplot for the given data in Example 3.15 in Python, you can use the `matplotlib` library. Here's the Python code that will generate a boxplot for the monthly starting salaries provided in the table:

```
import matplotlib.pyplot as plt

# Monthly starting salaries
salaries = [5850, 5950, 6050, 5880, 5755, 5710, 5890, 6130,
    5940, 6325, 5920, 5880]

# Create a boxplot
plt.figure(figsize=(10, 6))
plt.boxplot(salaries, vert=False, patch_artist=True,
            boxprops=dict(facecolor='lightblue', color='blue
    '),
            whiskerprops=dict(color='blue'),
            capprops=dict(color='blue'),
            medianprops=dict(color='red'))

# Add titles and labels
plt.title('Boxplot of Monthly Starting Salaries for Business
     School Graduates')
plt.xlabel('Monthly Starting Salary ($)')
plt.grid(True)

# Show plot
plt.show()
```

3.5.11 Exercises

1. The dataset represents the heights (in cm) of 20 plants measured over a period:

   $$40, 42, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63$$

   Compute the 1st decile (D1), 5th decile (D5, which is the median), and the 9th decile (D9) of the plant heights.

2. Given the following dataset of monthly rainfall (in mm) for 8 cities:

   $$100, 110, 120, 150, 170, 190, 200, 220$$

   (a) Compute the quartiles (Q1, Q2, Q3) for the dataset.

   (b) Detect any potential outliers using the IQR method.

3. The following dataset represents the scores of 25 students in an exam:

   $$45, 47, 48, 50, 51, 53, 55, 57, 58, 60, 62, 63, 65,$$

   $$67, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88$$

   (a) Calculate the 10th percentile (P10) and the 80th percentile (P80) of the exam scores.

DRAFT

(b) Detect any potential outliers using the IQR method.

4. The following dataset represents the number of hours spent on the internet per week by a sample of 20 people:

$$5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24$$

(a) Compute the quartiles (Q1, Q2, Q3) for the dataset.

(b) Find the 25th percentile (P25) and the 75th percentile (P75) of the dataset.

5. The following dataset represents housing prices (in thousands of dollars) in a neighborhood:

$$220, 230, 250, 270, 290, 310, 330, 350, 370, 400$$

(a) Compute the five-number summary.

(b) Calculate the interquartile range (IQR) of the housing prices.

(c) Draw a boxplot and comment on your findings.

6. The following dataset represents the test scores of 15 students:

$$55, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86$$

(a) Compute the five-number summary.

(b) Find the 25th percentile and the 90th percentile of the test scores.

(c) Draw a boxplot and comment on your findings.

7. The dataset below represents the number of goals scored by a soccer team over 10 games:

$$1, 2, 3, 3, 4, 5, 6, 6, 7, 8$$

(a) Calculate the five-number summary.

(b) Draw a boxplot based on the five-number summary.

(c) Label the components of the boxplot, including the minimum, Q1, median, Q3, maximum, and any potential outliers.

8. Consider the dataset representing the heights (in cm) of 20 individuals:

$$160, 165, 170, 175, 180, 185, 190, 195, 200, 205,$$

$$210, 215, 220, 225, 230, 235, 240, 245, 250, 255$$

(a) Compute the five-number summary for the dataset.

(b) Draw a boxplot and discuss the importance of the boxplot in visualizing the spread and identifying outliers.

DRAFT

9. Analyze the following dataset representing the number of hours spent on homework per week by 15 students:

$$5, 6, 7, 8, 8, 9, 10, 10, 11, 12, 12, 13, 14, 15, 20$$

   (a) Create a boxplot for this dataset.

   (b) Identify and describe the distribution of the data, including any potential outliers.

10. The following dataset represents the weights (in kg) of 18 animals:

$$2, 3, 3, 4, 5, 5, 6, 6, 7, 8, 8, 9, 10, 10, 11, 12, 13, 15$$

   (a) Use Python to compute the five-number summary and create a boxplot for this dataset.

   (b) Write a brief explanation of how Python can be used to visualize data distributions using boxplots.

3.6 Concluding Remarks

In conclusion, this chapter has provided a comprehensive overview of key numerical measures used in data exploration. By understanding and applying measures of central tendency, dispersion, and distribution shape, we equip ourselves with the tools needed to analyze and interpret data effectively. Each measure offers unique insights and, when used in combination, provides a robust understanding of the dataset's characteristics.

The exploration of quartiles, percentiles, deciles, and outlier detection further enhances our ability to segment and scrutinize data. The five-number summary and boxplots serve as powerful visual tools for summarizing and understanding data distributions. Mastery of these concepts, along with the accompanying Python code examples, will significantly contribute to your data science skills and enhance your ability to perform thorough data analysis.

We encourage you to apply these techniques to various datasets and practice interpreting the results to gain a deeper understanding of their implications. The exercises at the end of this chapter are designed to reinforce your learning and provide hands-on experience with these important data exploration tools.

3.7 Chapter Exercises

1. A physician collected an initial measurement of hemoglobin (g/L) after the admission of 10 inpatients to a hospital's department of cardiology. The hemoglobin measurements were $139, 158, 120, 112, 122, 132, 97, 104, 159, 129$.

DRAFT

(a) Calculate the mean and median hemoglobin levels.

(b) Find the range of the hemoglobin measurements.

(c) Compute the variance and standard deviation of the hemoglobin levels.

2. The average number of patients visiting a hospital each day over a week is recorded as $150, 160, 170, 140, 155, 165, 160$.

(a) What is the mean number of patients per day?

(b) Calculate the variance and standard deviation of the number of patients.

(c) Find the range of the daily patient counts.

3. The average score of 20 students on a test is 78, and the variance of their scores is 36. Four new students with scores 85, 90, 75, and 80 join the group.

(a) Compute the new average score for the group.

(b) Find the new variance and standard deviation of the test scores.

(a) Compute the new average score for the group.

(b) Find the variance and standard deviation of the test scores after adding the new students.

4. Compare the arithmetic means of the following two datasets:

- Dataset A: $55, 60, 65, 70, 75$
- Dataset B: $50, 60, 70, 80, 90$

(a) Compare the means of Dataset A and Dataset B.

(b) Compute the range, variance, and standard deviation for Dataset A.

(c) Compute the range, variance, and standard deviation for Dataset B.

5. If a class has two sections with 25 and 30 students, and the average scores for the sections are 85 and 90, respectively, find the weighted average score for the entire class.

(a) Find the weighted average score for the entire class.

(b) Calculate the variance and standard deviation of the scores, assuming that the scores in each section are the same.

6. A researcher collects the following data on the number of new cases of a disease per month: $20, 25, 30, 35, 40, 45, 50$.

DRAFT

(a) What is the median number of new cases?

(b) Compute the mean number of new cases.

(c) Find the range, variance, and standard deviation of the number of new cases.

7. For the dataset of daily temperatures: $22, 24, 21, 25, 23, 26, 24, 27$,

   (a) Calculate the median temperature.

   (b) Find the mean, range, variance, and standard deviation of the daily temperatures.

8. Calculate the medians of the following datasets:

   - Dataset A: $5, 8, 7, 6, 10$
   - Dataset B: $3, 6, 5, 7, 8, 9$

   (a) Find the median of Dataset A and Dataset B.

   (b) For Dataset A and Dataset B, calculate the range, variance, and standard deviation.

9. In a study on the number of hours spent on exercise per week by a group of individuals, the hours are recorded as $4, 6, 5, 7, 9, 5, 6, 7$.

   (a) Find the median number of hours spent exercising per week.

   (b) Compute the mean, range, variance, and standard deviation of the hours spent on exercise.

10. A car travels at speeds of 50 km/h for the first part of the trip and 70 km/h for the second part. If the distances traveled are the same, what is the harmonic mean of the two speeds?

    (a) Calculate the harmonic mean of the two speeds.

    (b) Discuss how the harmonic mean is used to find the average rate in this context.

11. A biostatistics researcher records the growth rates of a certain plant species over three months as $1.2, 1.5, 1.8$.

    (a) What is the geometric mean growth rate?

    (b) Find the variance and standard deviation of the growth rates.

12. The annual returns of two investment portfolios over the past 3 years are $10\%, 15\%, 5\%$ and $7\%, 12\%, 8\%$.

    (a) Compute the geometric mean return for each portfolio.

DRAFT

(b) Compare the geometric means of the two portfolios and discuss their implications.

13. A pharmaceutical company wants to calculate the average effectiveness of a new drug across three different studies with effectiveness rates of $0.85, 0.90, 0.80$.

    (a) What is the geometric mean effectiveness?

    (b) Compute the variance and standard deviation of the effectiveness rates.

14. A nutritionist evaluates the average calorie intake of patients based on three different dietary plans. The average calorie intake (in kcal) per day for each plan and the number of patients on each plan are as follows:

    - Plan A: 2000 kcal/day with 15 patients
    - Plan B: 2500 kcal/day with 25 patients
    - Plan C: 2200 kcal/day with 10 patients

    Calculate the weighted mean calorie intake per day for all patients.

15. A professor is calculating the overall grade for a student based on the grades from three different assessments with the following weights:

    - Exam 1: 85 (weight 40%)
    - Exam 2: 90 (weight 30%)
    - Final Exam: 88 (weight 30%)

    Find the weighted mean grade for the student.

16. This exercise focuses on comparing error rates for two software projects. You will analyze the central tendency and variability of the error rates.

    - **Project Alpha:** 5, 7, 6, 8, 5, 9, 6, 7, 8, 5, 7, 6
    - **Project Beta:** 8, 10, 9, 7, 11, 9, 8, 10, 11, 9, 12, 10

    Using the above data, answer the following questions.

    (a) Draw a boxplot for the error rates for Project Alpha and Project Beta.

    (b) Which project has the lower median number of errors?

    (c) Compare the spread of errors in each project. Which project shows greater variability?

DRAFT

(d) Look for outliers and discuss what they might indicate about the quality of the code in each project.

(e) Which project appears to have better performance based on the boxplots?

17. This exercise involves analyzing BMI measurements from the start of a study and after six months. Compare the central tendency and variability of BMI over time.

- **Start of Study:** 22.5, 23.1, 22.8, 21.9, 23.5, 22.7, 23.0, 21.5, 22.3, 22.9, 21.8, 23.2
- **After 6 Months:** 21.0, 21.5, 21.8, 20.8, 22.0, 21.3, 21.6, 20.7, 21.1, 21.9, 20.9, 21.4

Using the above data, answer the following questions.

(a) Draw a boxplot for BMI at the start of the study and after six months.

(b) How did the median BMI change from the start to after six months?

(c) Did the diet intervention lead to a reduction in the variability of BMI?

(d) Are there any outliers in the BMI data at the start or after six months? What can be inferred from them?

(e) Based on the boxplots, evaluate the effectiveness of the diet intervention.

18. This exercise involves analyzing cholesterol levels across three different groups. You will interpret the boxplots to compare the cholesterol levels between these groups.

- **Group A:** 190, 200, 195, 210, 205, 215, 202, 198, 220, 210, 195, 200
- **Group B:** 180, 185, 190, 175, 195, 190, 180, 185, 175, 190, 185, 180
- **Group C:** 210, 220, 215, 230, 225, 240, 220, 235, 225, 230, 240, 215

Using the above data, answer the following questions.

(a) Draw a boxplot for the cholesterol levels for Groups A, B, and C.

(b) Which group has the highest median cholesterol level?

(c) Which group has the smallest interquartile range?

(d) Are there any outliers in each group? If so, which group has the most outliers?

(e) Compare the variability of cholesterol levels between the three groups.

DRAFT

Chapter 4

Introduction to Probability

4.1 Introduction

In the realm of data science, understanding probability is essential for making informed decisions based on uncertain and incomplete information. Probability theory provides the mathematical foundation for analyzing data, modeling uncertainty, and deriving insights from complex datasets. As data scientists, we frequently encounter situations where outcomes are not deterministic but rather subject to variability and chance. Probability offers tools and frameworks to quantify this uncertainty and to make predictions that guide decision-making.

At its core, probability is concerned with measuring the likelihood of various outcomes in uncertain situations. Whether it's predicting customer behavior, assessing risk, or evaluating the effectiveness of an algorithm, probability helps us to model and interpret the inherent randomness in data. By applying probability theory, we can develop robust statistical models, conduct rigorous hypothesis testing, and perform meaningful data analysis.

In this chapter, we will lay the groundwork for understanding probability within the context of data science. We will start with the basic concepts that form the foundation of probability theory, including experiments, sample spaces, and events. We will then explore different methods of assigning probabilities, such as classical, empirical, and subjective approaches, and examine how these methods apply to real-world data problems.

As we delve deeper, we will cover key topics such as joint and marginal probabilities, conditional probability, and posterior probabilities. Each of these concepts is crucial for analyzing relationships between variables, updating beliefs based on new data, and making predictions about future events.

By the end of this chapter, you will gain a solid understanding of proba-

DRAFT

bility and its applications in data science. This knowledge will equip you with the tools needed to tackle complex data challenges and to make data-driven decisions with confidence.

4.2 Basic Concepts

4.2.1 Experiment

In the context of probability, an **experiment** refers to any process or action that generates a set of outcomes. The experiment is conducted under specified conditions, and the outcomes of interest are observed and recorded. Each outcome of an experiment is uncertain, but over repeated trials, patterns emerge that allow us to assign probabilities to different outcomes.

Illustrative Example

To explore the concept of an **experiment**, consider the situation of assessing the fairness of a coin used in a cricket match to decide which team bats or bowls first. The procedure begins by repeatedly flipping the coin under similar conditions, for example, 100 times, and recording the outcome of each flip.

Each flip represents a single **trial** of the experiment, and the possible **outcomes** are either heads or tails. Since the result of each flip is uncertain, this process meets the criteria of a probabilistic experiment. After conducting all 100 flips, we analyze the results by calculating the relative frequencies of heads and tails. These observed frequencies are then compared to the theoretical probability of 0.5 for each outcome.

If the proportion of heads is close to 0.5, it suggests that the coin is likely fair. On the other hand, a significant deviation from 0.5 may indicate that the coin is biased. Through this experiment, we gather evidence to support or refute the fairness of the coin based on empirical data.

4.2.2 Random Experiment

A **random experiment** is a specific type of experiment where the outcome is subject to chance and cannot be predicted with certainty. The outcomes are uncertain and vary each time the experiment is performed.

Examples

- **Rolling a Die**: A random experiment with outcomes {1, 2, 3, 4, 5, 6}, where each outcome is unpredictable.
- **Flipping a Coin**: A random experiment with two possible outcomes: heads or tails.

DRAFT

- **Drawing a Card**: A random experiment from a deck of 52 cards, where each card is equally likely to be drawn.

4.2.3 Sample Space

In data science, the concept of a sample space is essential for understanding the possible outcomes of a random experiment or a data-generating process. The sample space is the set of all possible outcomes or values that a random variable can take.

Sample Space: The set of all possible outcomes of a random experiment is denoted as the sample space, typically represented by S or Ω.

Consider the experiment of rolling a fair six-sided die. The possible outcomes of this experiment are the numbers that appear on the top face of the die after a roll. Then the sample space S for this experiment is the set of all possible outcomes. That is,

$$S = \{1, 2, 3, 4, 5, 6\}$$

Here are some examples of sample spaces in different contexts within data science:

- The sample space for tossing a fair coin is

$$S = \{\text{Heads}, \text{Tails}\}.$$

- Sample Space for tossing two coins:

$$S = \{(\text{Heads, Heads}), (\text{Heads, Tails}), (\text{Tails, Heads}), (\text{Tails, Tails})\}.$$

- The sample space S for rolling two six-sided dice consists of all possible ordered pairs (x_1, x_2), where x_1 represents the outcome of the first die and x_2 represents the outcome of the second die. Since each die has 6 faces, the sample space contains 36 possible outcomes:

$$S = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6), \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6), \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6), \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6), \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6), \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}$$

This sample space shows all the possible outcomes when two dice are rolled simultaneously.

DRAFT

4.2.4 Events

Events are subsets of the sample space, representing specific outcomes or combinations of outcomes.

> **Event:** A subset of the sample space.

For example, consider the experiment of rolling a fair six-sided die. The event A of rolling an even number is:

$$A = \{2, 4, 6\}$$

4.3 Probability

Probability is a branch of mathematics that deals with the likelihood of different outcomes in uncertain situations. It quantifies the chance of an event occurring, providing a way to model and analyze randomness. In essence, probability helps us understand and predict the behavior of systems in which outcomes are not deterministic, but rather subject to chance.

> **Probability:** A numerical measure of the likelihood that a particular event will occur. It quantifies uncertainty by assigning a value between 0 and 1, where 0 means the event will not happen (impossible) and 1 means the event will definitely happen (certain).

We use the notation $P(A)$ or $\Pr(A)$ to represent the probability that event A occurs. Here, P is the probability function that assigns a value to an event based on how likely it is to happen.

Example: Rolling a Die

Consider the experiment of rolling a fair six-sided die. The sample space is:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let event A be "rolling an even number." The favorable outcomes are:

$$A = \{2, 4, 6\}$$

Since all outcomes are equally likely:

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} = \frac{3}{6} = 0.5$$

Therefore, the probability of rolling an even number is 0.5.

DRAFT

4.3.1 Types of Probability

Probability can be categorized into different types based on how it is determined or calculated. Below are the key types:

(a). Classical (Theoretical) Probability

(b). Experimental (Empirical) Probability

(c). Subjective Probability

(a). Classical (Theoretical) Probability

Classical or theoretical probability is based on the assumption that all outcomes of a random experiment are **equally likely**. This is often used when the outcomes of an experiment are known and finite. If an experiment has n equally likely outcomes and an event A consists of m of these outcomes, the probability of event A is given by:

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} = \frac{m}{n}.$$

Example 1: For a fair six-sided die, the total number of possible outcomes is 6 (since the die has six faces). If we are interested in the probability of rolling a 3, there is only one favorable outcome (rolling a 3). Therefore, the probability is:

$$P(\text{rolling a 3}) = \frac{1}{6}$$

Example 2: For a standard deck of 52 playing cards, the total number of possible outcomes is 52. If we want to know the probability of drawing an Ace, there are 4 Aces in the deck (one in each suit). Therefore, the probability is:

$$P(\text{drawing an Ace}) = \frac{4}{52} = \frac{1}{13}$$

Classical probability is particularly useful for situations where the outcomes are well-defined, and each outcome is equally likely, such as rolling dice, drawing cards, or selecting outcomes from a set of equally likely possibilities.

(b). Experimental (Empirical) Probability

The empirical approach, also known as the frequentist approach, assigns probabilities based on observed frequencies from experimental data. This approach is often used in practical situations where we have observed data. If an experiment is repeated N times and event A occurs n_A times, the probability of event A is estimated by the relative frequency:

DRAFT

$$P(A) = \frac{\text{Number of times the event } A \text{ occurred}}{\text{Total number of observations}} = \frac{n_A}{N}.$$

This method is particularly useful when it is difficult or impractical to calculate theoretical probabilities, or when we want to verify theoretical predictions by comparing them with actual outcomes. The empirical approach to probability relies on the principle known as the **law of large numbers**. This principle suggests that as the number of observations increases, the estimate of the probability becomes more accurate. Therefore, by gathering more data, one can obtain a more precise estimation of the probability.

Law of large numbers: As the number of trials or observations increases, the empirical probability of an event will get closer to its actual probability.

Example 1: For example, if we flip a coin 100 times and get 52 heads, the estimated probability of getting a head would be:

$$P(\text{Heads}) = \frac{\text{Number of heads}}{\text{Total number of flips}} = \frac{52}{100} = 0.52$$

Based on the empirical data, the estimated probability of getting heads on a coin flip is 0.52, or 52%.

Example 2: The air route between Dhaka and Chattogram is among the busiest in Bangladesh, with airlines aiming to maintain high punctuality. To assess performance, data from 100 monitored flights were analyzed:

- On-time (successful) flights: 95
- Delayed or canceled (unsuccessful) flights: 5

The empirical probability of a successful flight is:

$$P(\text{Success}) = \frac{95}{100} = 0.95$$

The empirical probability of an unsuccessful flight is:

$$P(\text{Failure}) = \frac{5}{100} = 0.05$$

Thus, based on observed data, there is a 95% chance of a flight being on time and a 5% chance of delay or cancellation.

(c). Subjective Approach

The subjective approach assigns probabilities based on **personal judgment** or **belief about the likelihood of an event**. This method does not rely on mathematical calculations or empirical data but rather on an individual's

DRAFT

intuition or experience. For an event A, the subjective probability is denoted as:

$$P(A) = \text{Subjective belief about } A$$

In this approach, probabilities are not necessarily based on frequency or equal likelihood but on personal estimation.

Example 1: Consider an entrepreneur deciding whether to launch a new product. Since there is no historical data or empirical studies available for this specific product, the entrepreneur uses their expertise and market knowledge to estimate the likelihood of success.

The entrepreneur assesses various factors:

- Knowledge of current market trends.
- Feedback from potential customers.
- Expertise of the team involved.
- Analysis of the competitive landscape.

Based on these factors, the entrepreneur might estimate the probability of the product's success to be 70%. This subjective estimate is derived from their personal judgment and experience rather than from data analysis.

$$P(\text{Success}) = 0.70$$

This subjective probability is based on the entrepreneur uses their expertise and market knowledge, rather than on empirical data or mathematical models.

Example 2: Consider a football game between Team A and Team B. Based on your personal judgment and knowledge of the teams, you estimate the probability of Team A winning the game.

Let's denote:

- $P(A)$ as the probability of Team A winning.
- $P(B)$ as the probability of Team B winning.

Based on your assessment, you estimate that:

$$P(A) = 0.70$$

This means you believe there is a 70% chance that Team A will win the game. This probability is derived from your subjective evaluation of the teams' recent performances, player conditions, and other relevant factors.

DRAFT

4.3.2 Properties of the Probability

A set of probability values for an experiment with a sample space

$$S = \{A_1, A_2, \ldots, A_n\}$$

consists of some probabilities $P(A_1), P(A_2), \ldots, P(A_n)$ that must satisfy

$$0 \leq P(A_1) \leq 1, \quad 0 \leq P(A_2) \leq 1, \quad \ldots, \quad 0 \leq P(A_n) \leq 1$$

and

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1.$$

Properties of the Probability:

(i). The probability of an event is always a number between 0 and 1.

(ii). The sum of the probabilities of all mutually exclusive events is always 1.

For example, when tossing a fair coin, the probability of getting heads is 0.5, and the probability of getting tails is also 0.5, which are both between 0 and 1. The sum of these probabilities is

$$0.5 + 0.5 = 1$$

showing that the total probability of all possible outcomes (heads or tails) is always 1.

Problem 4.1. *An experiment has five outcomes, I, II, III, IV, and V. If* $P(I) = 0.08$*,* $P(II) = 0.20$*, and* $P(III) = 0.33$*, (a) what are the possible values for the probability of outcome V? (b) If outcomes IV and V are equally likely, what are their probability values?*

Solution

(a).
An experiment has five outcomes: I, II, III, IV, and V. Given the probabilities for outcomes I, II, and III are:

$$P(\text{I}) = 0.08, \quad P(\text{II}) = 0.20, \quad P(\text{III}) = 0.33$$

We need to determine the possible values for the probability of outcome V. First, we find the sum of the given probabilities:

$$P(\text{I}) + P(\text{II}) + P(\text{III}) = 0.08 + 0.20 + 0.33 = 0.61$$

Since the sum of the probabilities of all outcomes must equal 1, the sum of the probabilities of outcomes IV and V is:

DRAFT

$$P(\text{IV}) + P(\text{V}) = 1 - 0.61 = 0.39$$

Thus, the possible values for the probability of outcome V, denoted as $P(\text{V})$, depend on the probability of outcome IV, denoted as $P(\text{IV})$:

$$P(\text{V}) = 0.39 - P(\text{IV})$$

Since, $0 \leq P(\text{IV}) \leq 0.39$, then the possible values for the probability of outcome V are

$$0 \leq P(\text{V}) \leq 0.39$$

(b).

If outcomes IV and V are equally likely, then:

$$P(\text{IV}) = P(\text{V})$$

Let $P(\text{IV}) = P(\text{V}) = x$. Then:

$$2x = 0.39 \quad \Rightarrow \quad x = \frac{0.39}{2} = 0.195$$

Therefore, the probabilities for outcomes IV and V are:

$$P(\text{IV}) = P(\text{V}) = 0.195$$

Problem 4.2. *An experiment has three outcomes, I, II, and III. If outcome I is twice as likely as outcome II, and outcome II is three times as likely as outcome III, what are the probability values of the three outcomes?*

Solution

An experiment has three outcomes: I, II, and III. Let the probabilities of these outcomes be $P(\text{I})$, $P(\text{II})$, and $P(\text{III})$, respectively.

Let $P(\text{III}) = x$. Then:

- $P(\text{II}) = 3x$ (since outcome II is three times as likely as outcome III).
- $P(\text{I}) = 2 \cdot P(\text{II}) = 2 \cdot 3x = 6x$ (since outcome I is twice as likely as outcome II).

Since the sum of the probabilities of all outcomes must equal 1, we have:

$$P(\text{I}) + P(\text{II}) + P(\text{III}) = 1$$

Substituting the values, we get:

$$6x + 3x + x = 1$$

DRAFT

or,

$$10x = 1$$

$$\therefore x = \frac{1}{10} = 0.1$$

Thus, the probabilities of the three outcomes are:

$$P(\text{III}) = x = 0.1$$
$$P(\text{II}) = 3x = 3 \times 0.1 = 0.3$$
$$P(\text{I}) = 6x = 6 \times 0.1 = 0.6$$

Therefore, the probability values of the three outcomes are:

$$P(\text{I}) = 0.6, \quad P(\text{II}) = 0.3, \quad P(\text{III}) = 0.1$$

4.3.3 Union of Events

The union of two events A and B, denoted $A \cup B$, represents the event that either A, B, or both occur. Formally:

$$A \cup B = \{\omega \mid \omega \in A \text{ or } \omega \in B\}$$

Example: Consider rolling a standard six-sided die. Let:

- A be the event "rolling an even number" (i.e., $A = \{2, 4, 6)\}$)
- B be the event "rolling a number greater than 3" (i.e., $B = \{4, 5, 6)\}$)

The union $A \cup B$ represents rolling a number that is either even or greater than 3 (or both). The possible outcomes for

$$A \cup B = \{2, 4, 5, 6\}.$$

The probability of $A \cup B$ is given by:

$$P(A \cup B) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{4}{6} = \frac{2}{3}$$

4.3.4 Intersection of Events

The intersection of two events A and B, denoted $A \cap B$, represents the event that both A and B occur simultaneously. Formally:

$$A \cap B = \{\omega \mid \omega \in A \text{ and } \omega \in B\}$$

Example: Using the same die roll, let:

- A be the event "rolling an even number" (i.e., $A = \{2, 4, 6)\}$)

DRAFT

- B be the event "rolling a number greater than 3" (i.e., $B = \{4, 5, 6)\}$)

The intersection $A \cap B$ represents rolling a number that is both even and greater than 3.

The possible outcomes for

$$A \cap B = \{4, 6\}.$$

The probability of $A \cap B$ is given by:

$$P(A \cap B) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{2}{6} = \frac{1}{3}$$

4.3.5 Complementary Event

A **complementary event** of an event A, denoted by A^c or $\overline{A}$ or A', consists of all outcomes in the sample space S that are not in A.
Example: Consider a six-sided die.

- **Event A:** Rolling an even number:

$$A = \{2, 4, 6\}$$

- **Complementary Event A^c:** Rolling an odd number:

$$A^c = \{1, 3, 5\}$$

The probability of the complementary event A^c is given by:

$$P(A^c) = 1 - P(A)$$

Example: If A is the event of getting a head when flipping a coin, then the complementary event A^c is the event of getting a tail. If $P(A) = 0.5$, then $P(A^c) = 1 - 0.5 = 0.5$.

4.3.6 Odds and Odds Ratio

Odds: In probability theory, **odds** provide an alternative way to express the likelihood of an event. While probability measures how likely an event is to occur out of all possible outcomes, odds compare the chance of the event happening to the chance of it not happening. If A is an event with probability $P(A)$, then the *odds in favor* of A are defined as the ratio of the probability that the event occurs to the probability that it does not occur:

$$\text{Odds in favor of } A = \frac{P(A)}{1 - P(A)}.$$

DRAFT

Interpretation

- If the odds in favor of an event are greater than 1, it means the event is more likely to occur than not.
- If the odds are less than 1, the event is less likely to occur.
- If the odds equal 1, the event is equally likely to occur or not occur.

Odds: The *odds* in favor of an event A are defined as the ratio of the probability that the event A occurs to the probability that the event A does not occur (i.e., the complement of A). Mathematically, the odds in favor of A are given by:

$$\text{Odds in favor of } A = \frac{P(A)}{P(A^c)}$$

where $P(A^c)$ is the probability of the complement of A.

Example: Suppose the probability that a team wins a match is 0.75. Then, the odds in favor of the team winning are:

$$\text{Odds in favor of win} = \frac{0.75}{1-0.75} = \frac{0.75}{0.25} = 3,$$

which means the team is three times more likely to win than to lose.

Odds Ratio: The **odds ratio (OR)** is a measure used to compare the odds of an event occurring in two different groups. It is defined as the ratio of the odds of the event in one group to the odds in another group.

Suppose we have two groups:

- Group 1: probability of event A_1 is P_1
- Group 2: probability of event A_2 is P_2

Then the odds ratio is:

$$\text{Odds Ratio} = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\dfrac{P_1}{1-P_1}}{\dfrac{P_2}{1-P_2}} = \frac{P_1(1-P_2)}{P_2(1-P_1)}$$

Interpretation

- OR $= 1$: The odds of the event are the same in both groups.
- OR > 1: The event is more likely in Group 1 than in Group 2.
- OR < 1: The event is less likely in Group 1 than in Group 2.

DRAFT

Example: (Comparing Two Treatments) A study compares two treatments for a medical condition:

- In Group 1 (new treatment), 80 out of 100 patients improved: $P_1 = 0.80$
- In Group 2 (standard treatment), 60 out of 100 patients improved: $P_2 = 0.60$

Compute the odds for each group:

$$\text{Odds}_1 = \frac{0.80}{1 - 0.80} = \frac{0.80}{0.20} = 4$$

$$\text{Odds}_2 = \frac{0.60}{1 - 0.60} = \frac{0.60}{0.40} = 1.5$$

Now compute the odds ratio:

$$\text{Odds Ratio} = \frac{4}{1.5} = \frac{8}{3} \approx 2.67$$

Interpretation: The odds of improvement are approximately 2.67 times higher with the new treatment compared to the standard treatment.

Problem 4.3. *Suppose p is the probability of the success.*

(a) If the odds is 1, what is p?

(b) If the odds is 2, what is p?

(c) If $p = 0.25$, what is the odds?

Solution

(a). The odds in favor of success are given by:

$$\text{Odds} = \frac{p}{1 - p}$$

If the odds are 1, then:

$$1 = \frac{p}{1 - p}$$

Solving for p:

$$1 - p = p \quad \Rightarrow \quad 1 = 2p \quad \Rightarrow \quad p = \frac{1}{2}$$

So, $p = 0.5$.

DRAFT

(b). If the odds are 2, then:

$$2 = \frac{p}{1-p}$$

Solving for p:

$$2(1-p) = p \quad \Rightarrow \quad 2 - 2p = p \quad \Rightarrow \quad 2 = 3p \quad \Rightarrow \quad p = \frac{2}{3}$$

So, $p = \frac{2}{3}$.

(c). If $p = 0.25$, then the odds are:

$$\text{Odds} = \frac{p}{1-p} = \frac{0.25}{1-0.25} = \frac{0.25}{0.75} = \frac{1}{3}$$

So, the odds are $\frac{1}{3}$.

4.3.7 Equally Likely Events

In probability theory, equally likely events are events that have the same probability of occurring. If all outcomes in the sample space are equally likely, then the probability of any specific event can be calculated by dividing the number of favorable outcomes by the total number of possible outcomes.

Equally Likely Events: Events that have the same probability of occurring.

Example: Consider the experiment of rolling a fair six-sided die. The sample space is:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Since the die is fair, each of the six outcomes is equally likely. The probability of each outcome is:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}.$$

4.3.8 Mutually Exclusive Events

In probability theory, mutually exclusive events are events that cannot happen at the same time. In other words, if one event occurs, the other cannot occur at the same time.

Mutually Exclusive Events: Two events A and B are said to be mutually exclusive (or disjoint) if they cannot occur at the same time. That is, A and B are mutually exclusive if

$$A \cap B = \emptyset$$

DRAFT

where $\cap$ denotes the **intersection** of events, and $\emptyset$ represents the empty set, indicating that there are no outcomes common to both A and B.

Example:

Consider rolling a standard six-sided die. Let:

- A be the event "rolling a 2"
- B be the event "rolling a 5"

The events A and B are mutually exclusive because you cannot roll a 2 and a 5 at the same time.

Additivity for Mutually Exclusive Events

If A and B are mutually exclusive events, then the probability of their union is the sum of their individual probabilities:

$$P(A \cup B) = P(A) + P(B)$$

Generalization

For any finite or countable collection of mutually exclusive events $A_1, A_2, \ldots, A_n$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

4.3.9 Probability Axioms

The probability of an event A_i; $i = 1, 2, \ldots, n$, denoted as $P(A_i)$, satisfies the following axioms:

(i). $0 \leq P(A_i) \leq 1$,

(ii). $\sum_{i=1}^{n} P(A_i) = 1$,

(iii). For any sequence of **mutually exclusive events** $\{A_i\}$, we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

For mutually exclusive events $A, B \in S$, we have

$$P(A \cup B) = P(A) + P(B)$$

This is an ***addition rule of probability for two mutually exclusive events***

DRAFT

If A and B are not mutually exclusive, then the addition rule is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

If two events are mutually exclusive, then the probability of both occurring is denoted as $P(A \cap B)$ and

$$P(A \text{ and } B) = P(A \cap B) = 0.$$

Problem 4.4. *A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?*

Solution

- $P(2) = \frac{1}{6}$ and $P(5) = \frac{1}{6}$
- Therefore,

$$\begin{aligned} P(2 \text{ or } 5) = P(2 \cup 5) &= P(2) + P(5) \\ &= \frac{1}{6} + \frac{1}{6} \\ &= \frac{2}{6} \\ &= \frac{1}{3} \end{aligned}$$

Problem 4.5. *In a Math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade. If a student is chosen at random from the class, what is the probability of choosing a girl or an A-grade student?*

Solution

- $P(\text{girl}) = \frac{13}{30}$, $P(A\text{-grade student}) = \frac{9}{30}$ and $P(\text{girl} \cap A\text{-grade student}) = \frac{5}{30}$
- Therefore,

$$\begin{aligned} P(\text{girl or } A\text{-grade student}) &= P(\text{girl}) + P(A\text{-grade student}) \\ &\quad - P(\text{girl} \cap A\text{-grade student}) \\ &= \frac{13}{30} + \frac{9}{30} - \frac{5}{30} \\ &= \frac{17}{30} \end{aligned}$$

Problem 4.6. *Suppose two dice are thrown together. What is the probability that at least one 6 is obtained?*

DRAFT

Solution

Since each die has 6 faces, the sample space contains $6 \times 6 = 36$ possible outcomes:

$$S = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6), \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6), \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6), \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6), \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6), \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}$$

Define the following events:

- A: The first die shows a 6
- B: The second die shows a 6

We are interested in $P(A \cup B)$, the probability that **at least one die shows a 6**.

First, compute each component:

$$P(A) = \frac{6}{36} = \frac{1}{6} \quad \text{(outcomes: } (6,1), (6,2), \ldots, (6,6)\text{)}$$

$$P(B) = \frac{6}{36} = \frac{1}{6} \quad \text{(outcomes: } (1,6), (2,6), \ldots, (6,6)\text{)}$$

$$P(A \cap B) = \frac{1}{36} \quad \text{(outcome: } (6,6)\text{)}$$

Now apply the formula for the probability of the union:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$$

Thus, the probability that at least one 6 appears is

$$P(\text{at least one 6}) = \frac{11}{36}.$$

4.4 Joint and Marginal Probabilities

In probability theory, it is often important to analyze the likelihood of the occurrence of two or more events together, as well as the probabilities of individual events irrespective of others. This leads to the concepts of **joint probabilities** and **marginal probabilities**.

DRAFT

4.4.1 Joint Probability of Events

The **joint probability** of two events A and B, denoted by $P(A \cap B)$, is the probability that both events happen simultaneously. For example, if A is "being a smoker" and B is "having a myocardial infarction (heart attack)," then $P(A \cap B)$ represents the probability that a person is both a smoker and has experienced a myocardial infarction.

4.4.2 Marginal Probability of Events

The **marginal probability** of an event is the probability that the event occurs without consideration of any other event. For event A, this is denoted $P(A)$.

Example: (Smoking Status and Myocardial Infarction) Suppose a medical study surveyed 1,000 adults and recorded their smoking status and whether they had suffered a myocardial infarction (MI). The data is summarized below:

MI Status	Smoker	Non-Smoker	Total
Had MI	90	60	150
No MI	310	540	850
Total	400	600	1000

From this data, the probabilities are:

- Joint probability that a person is a smoker and had a myocardial infarction:
$$P(\text{Smoker} \cap \text{MI}) = \frac{90}{1000} = 0.09.$$
- Joint probability that a person is a non-smoker and had no myocardial infarction:
$$P(\text{Non-Smoker} \cap \text{No MI}) = \frac{540}{1000} = 0.54.$$
- Marginal probability that a person is a smoker:
$$P(\text{Smoker}) = \frac{400}{1000} = 0.40.$$
- Marginal probability that a person had a myocardial infarction:
$$P(\text{MI}) = \frac{150}{1000} = 0.15.$$

DRAFT

The joint probabilities provide insight into how smoking status and myocardial infarction co-occur, while marginal probabilities summarize the overall likelihood of smoking or myocardial infarction individually. This analysis can be extended further to investigate associations or dependence between these health-related events.

Remark 4.4.1. *(Computing Probabilities from a Contingency Table) A **contingency table** summarizes the frequency distribution of two categorical variables. Each cell shows the count of observations corresponding to a specific combination of categories. Such tables are commonly used to compute both **joint** and **marginal** probabilities.*

Consider two categorical variables:

- ***Variable A** with categories A_1 and A_2*
- ***Variable B** with categories B_1 and B_2*

The contingency table in Table 4.1 summarizes the joint distribution of these two variables:

Table 4.1: Contingency Table

	A_1	A_2	**Total**
B_1	a	b	$a+b$
B_2	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

*The corresponding **joint probability table** is obtained by dividing each cell frequency by the total number of observations n. Each entry then represents the probability of observing the corresponding combination of categories.*

$$P(A_i \cap B_j) = \frac{\textit{Count in cell}}{n} \quad ; \quad i,j = 1,2.$$

The joint probability table is shown in Table 4.2:

Table 4.2: Joint Probability Table

	A_1	A_2	**Total**
B_1	$\frac{a}{n}$	$\frac{b}{n}$	$\frac{a+b}{n}$
B_2	$\frac{c}{n}$	$\frac{d}{n}$	$\frac{c+d}{n}$
Total	$\frac{a+c}{n}$	$\frac{b+d}{n}$	1

*Here, the row and column totals represent the **marginal probabilities**, and the individual cells represent the **joint probabilities** of specific combinations of categories of the two variables.*

DRAFT

4.5 Conditional Probability

Conditional probability is essential in data science because it helps model and understand how the probability of an event changes based on the occurrence of another event. It underpins Bayesian inference, supports feature engineering, enhances risk assessment, informs decision-making, aids in anomaly detection, and is pivotal in natural language processing tasks. This ability to adjust probabilities with new information is crucial for accurate predictions and data-driven insights.

Conditional Probability: Conditional probability is the probability of an event occurring given that another event has already occurred. The conditional probability of an event A given that event B has occurred is denoted by $P(A|B)$ and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0$$

One simple example of conditional probability concerns the situation in which two events A and B are mutually exclusive. Since mutually exclusive events have no common outcomes, the occurrence of event B makes the occurrence of event A impossible. Thus, intuitively, the probability of event A given that B has occurred should be zero. This is confirmed by the formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{P(B)} = 0.$$

Another example involves a scenario where event B is a subset of event A, denoted $B \subseteq A$. In this case, if event B occurs, event A must also occur. Thus, the probability of event A given that B has occurred should be one. This is supported by the formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Problem 4.7. *If somebody rolls a fair die without showing you but announces that the result is even, then what is the probability of scoring a 6?*

Solution

The sample space for a fair die roll is $S = \{1, 2, 3, 4, 5, 6\}$. The event (B) that the result is even is $B = \{2, 4, 6\}$.

$$P(B) = \frac{3}{6} = \frac{1}{2}$$

The event (A) of scoring a 6 given that the result is even is $A = \{6\}$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

DRAFT

Problem 4.8. *Suppose somebody rolls a red die and a blue die together without showing you, but announces that at least one 6 has been showed. What is the probability that the red die showed a 6?*

Solution

In the sample space S mentioned in Problem 4.6, we see the possible outcomes with at least one 6 are

$$B = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(1,6),(2,6),(3,6),(4,6),(5,6)\}$$

. Therefore, the number of outcomes with at least one 6 is 11. The number of outcomes where the red die scores a 6 is 6. That is,

$$A = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.$$

Hence, the conditional probability is

$$P(A|B) = \frac{6/36}{11/36} = \frac{6}{11}.$$

Problem 4.9. *Suppose somebody rolls a red die and a blue die together. What is the probability that the red die scores a 6 given that exactly one 6 of the two outcomes has been scored?*

Solution

In the sample space S mentioned in Problem 4.6, we see the possible outcomes with at least one 6 are

$$B = \{(6,1),(6,2),(6,3),(6,4),(6,5),(1,6),(2,6),(3,6),(4,6),(5,6)\}.$$

Therefore, the number of outcomes with exactly one 6 is 10 (i.e., excluding $(6,6)$). The number of outcomes where the red die scores a 6 and the blue die does not is 5 (i.e., $A = \{(6,1),(6,2),(6,3),(6,4),(6,5)\}$).

$$P(A|B) = \frac{5/36}{10/36} = \frac{1}{2}$$

Problem 4.10. *Consider the situation of the promotion status of* ***Male*** *and* ***Female*** *officers of a major metropolitan police force in the eastern United States. The force consists of 1200 officers, 960* ***Male*** *and 240* ***Female****. Over the past two years, 324 officers on the police force received promotions. After reviewing the promotion record, a committee of* ***Female*** *officers raised a discrimination case on the basis that 288* ***Male*** *officers had received promotions, but only 36* ***Female*** *officers had received promotions.*

DRAFT

	Male	*Female*	*Total*
Promoted	288	36	324
Not Promoted	672	204	876
Total	960	240	1200

*(i). Develop a joint probability table for these data. What are the marginal probabilities? Suppose a **Male** officer is selected randomly, what is the chance that the officer will be promoted?*

*(ii). Suppose a **Female** officer is selected randomly, what is the chance that the officer will not be promoted? Suppose an officer is selected randomly who got promoted, what is the chance that the officer will be **Male**?*

*(iii). Suppose an officer is selected randomly who did not get promotion, what is the chance that the officer will be **Female**?*

Solution

(i) Joint Probability Table and Marginal Probabilities

To develop the joint probability table, we divide each cell count by the total number of officers, which is 1200.

Joint Probability Table :

	Male	Female	Total
Promoted	$\frac{288}{1200}$	$\frac{36}{1200}$	$\frac{324}{1200}$
Not Promoted	$\frac{672}{1200}$	$\frac{204}{1200}$	$\frac{876}{1200}$
Total	$\frac{960}{1200}$	$\frac{240}{1200}$	1

Simplifying the fractions, we get:

	Male	Female	Total
Promoted	0.24	0.03	0.27
Not Promoted	0.56	0.17	0.73
Total	0.80	0.20	1

Marginal Probabilities:

- Probability of promotion: $\frac{324}{1200} = 0.27$
- Probability of not promotion: $\frac{876}{1200} = 0.73$

DRAFT

- Probability of being Male: $\frac{960}{1200} = 0.80$
- Probability of being Female: $\frac{240}{1200} = 0.20$

Probability that a randomly selected Male officer is promoted:

$$P(\text{Promoted} \mid \text{Male}) = \frac{288/1200}{960/1200} = 0.30$$

(ii) Probabilities for Female Officers and Promotion

Probability that a randomly selected Female officer is not promoted:

$$P(\text{Not Promoted} \mid \text{Female}) = \frac{204/1200}{240/1200} = 0.85$$

Probability that a randomly selected officer who got promoted is Male:

$$P(\text{Male} \mid \text{Promoted}) = \frac{288/1200}{324/1200} \approx 0.889$$

(iii) Probability for Officers Not Promoted

Probability that a randomly selected officer who did not get promoted is Female:

$$P(\text{Female} \mid \text{Not Promoted}) = \frac{204/1200}{876/1200} \approx 0.233$$

Multiplication Rule of Probability: The Multiplication Rule is used to find the probability that both events occur. For any two events A and B:

$$P(A \cap B) = P(A)\, P(B \mid A)$$

where,

- $P(A)$: probability that event A occurs
- $P(B \mid A)$: probability that event B occurs given that A has already occurred

4.6 Independent Events

In data science, understanding whether two events are independent is crucial for building models, especially those that rely on simplifying assumptions like Naive Bayes classifiers or probabilistic graphical models. This section explores what it means for events to be independent and how to test for independence

DRAFT

in real-world data.

Two events A and B are said to be **independent** if the occurrence of one event does not affect the probability of the other event occurring. In mathematical terms, this is expressed as:

$$P(A \cap B) = P(A) \cdot P(B).$$

For independent events, the following also holds true:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

and

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B).$$

These equations indicate that knowing the occurrence of one event does not change the probability of the other event.

Independent Events: Two events A and B are said to be **independent** if:

$$P(A \cap B) = P(A) \cdot P(B).$$

This equation implies

$$P(A \mid B) = P(A), \qquad P(B \mid A) = P(B).$$

Any one of these three conditions implies the other two.

Independence vs. Mutual Exclusivity

A common confusion arises between **independent** and **mutually exclusive** events. **Mutually exclusive** events cannot occur at the same time. If A and B are mutually exclusive:

$$P(A \cap B) = 0$$

However, unless $P(A) = 0$ or $P(B) = 0$, this violates the definition of independence:

$$P(A) \cdot P(B) > 0 \neq P(A \cap B)$$

Therefore, mutually exclusive events are *not* independent unless one of them has zero probability.

DRAFT

Example 1: Rolling a Die and Flipping a Coin

Consider rolling a fair six-sided die and flipping a fair coin. Let:

- A be the event that the die shows a 3.
- B be the event that the coin lands on heads.

The die has six possible outcomes: 1, 2, 3, 4, 5, or 6. Therefore, the sample space for the die roll is:

$$S_A = \{1, 2, 3, 4, 5, 6\}$$

The coin has two possible outcomes: heads (H) or tails (T). Therefore, the sample space for the coin flip is:

$$S_B = \{H, T\}$$

The outcome of rolling the die does not affect the outcome of flipping the coin, and vice versa. Therefore, events A and B are independent. We can verify this as follows:

$$P(A) = \frac{1}{6}, \quad P(B) = \frac{1}{2}$$

The combined sample space consists of 12 outcomes.

$$S = \begin{Bmatrix} (1,H), & (2,H), & (3,H), & (4,H), & (5,H), & (6,H), \\ (1,T), & (2,T), & (3,T), & (4,T), & (5,T), & (6,T) \end{Bmatrix}$$

$$P(A \cap B) = P(\text{die shows 3 and coin lands on heads}) = \frac{1}{12}$$

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Thus, the events are independent.

Example 2: Drawing Cards from a Deck Without Replacement

Consider drawing two cards from a standard deck of 52 cards without replacement. Let:

- A be the event that the first card drawn is a heart.
- B be the event that the second card drawn is a heart.

In this case, the events are not independent because the outcome of the first draw affects the probability of the second draw. If the first card is a heart, there are now only 12 hearts left in a deck of 51 cards, so:

$$P(A) = \frac{13}{52}, \quad P(B \mid A) = \frac{12}{51}$$

DRAFT

The probability of B if A occurs is different from $P(B)$ without conditioning on A, which is:

$$P(B) = \frac{13}{52}$$

Thus, A and B are not independent.

Problem 4.11. *A system has four computers. Computer 1 works with a probability of 0.88; computer 2 works with a probability of 0.78; computer 3 works with a probability of 0.92; computer 4 works with a probability of 0.85. Suppose that the operations of the computers are independent of each other.*

(a). Suppose that the system works only when all four computers are working. What is the probability that the system works?

(b). Suppose that the system works only if at least one computer is working. What is the probability that the system works?

(c). Suppose that the system works only if at least three computers are working. What is the probability that the system works?

Solution

(a). To find the probability in this scenario, we multiply the probabilities of all four computers working, as they are independent.

$$P(\text{system works}) = 0.88 \times 0.78 \times 0.92 \times 0.85 = 0.537$$

(b). To find the probability in this scenario, we find the complement of the probability that none of the computers are working. Then, the probability that at least one computer is working is the complement of the probability that none of the computers are working.

$$\begin{aligned} P(\text{system works}) &= 1 - P(\text{no computers working}) \\ &= 1 - ((1 - 0.88) \times (1 - 0.78) \times (1 - 0.92) \times (1 - 0.85)) \\ &= 0.9997 \end{aligned}$$

DRAFT

(c).

$$\begin{aligned}
P(\text{system works}) &= P(\text{all computers working}) \\
&\quad + P(\text{computers 1,2,3 working, computer 4 not working}) \\
&\quad + P(\text{computers 1,2,4 working, computer 3 not working}) \\
&\quad + P(\text{computers 1,3,4 working, computer 2 not working}) \\
&\quad + P(\text{computers 2,3,4 working, computer 1 not working}) \\
&= 0.537 + (0.88 \times 0.78 \times 0.92 \times (1 - 0.85)) \\
&\quad + (0.88 \times 0.78 \times (1 - 0.92) \times 0.85) \\
&\quad + (0.88 \times (1 - 0.78) \times 0.92 \times 0.85) \\
&\quad + ((1 - 0.88) \times 0.78 \times 0.92 \times 0.85) \\
&= 0.903
\end{aligned}$$

□

Problem 4.12. *Suppose that somebody secretly rolls two fair six-sided dice, and what is the probability that the face-up value of the first one is 2, given the information that their sum is no greater than 5?*

Solution

To find the probability that the face-up value of the first die is 2 given that the sum of the two dice is no greater than 5, we use the concept of conditional probability.

Let A be the event that the face-up value of the first die is 2, and B be the event that the sum of the two dice is no greater than 5. We want to find $P(A \mid B)$.

The conditional probability $P(A \mid B)$ is given by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

First, we determine $P(B)$. The possible outcomes for the sum of the two dice being no greater than 5 are:

$$\begin{aligned}
&(1,1), (1,2), (1,3), (1,4), \\
&(2,1), (2,2), (2,3), \\
&(3,1), (3,2), \\
&(4,1)
\end{aligned}$$

There are 10 such outcomes, and since there are 36 possible outcomes when rolling two dice, the probability $P(B)$ is:

$$P(B) = \frac{10}{36} = \frac{5}{18}$$

DRAFT

Next, we determine $P(A \cap B)$, which is the probability that the first die is 2 and the sum of the dice is no greater than 5. The possible outcomes for this are:

$$(2,1), (2,2), (2,3)$$

There are 3 such outcomes, so the probability $P(A \cap B)$ is:

$$P(A \cap B) = \frac{3}{36} = \frac{1}{12}$$

Now we can calculate $P(A \mid B)$:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{12}}{\frac{5}{18}} = \frac{1}{12} \times \frac{18}{5} = \frac{18}{60} = \frac{3}{10}$$

Thus, the probability that the face-up value of the first die is 2 given that their sum is no greater than 5 is $\frac{3}{10}$. □

4.7 Posterior Probabilities

4.7.1 Law of Total Probability

Consider a sample space S partitioned into mutually exclusive events $A_1, A_2, \ldots, A_k$. This means, the sample space S can be written as

$$S = A_1 \cup A_2 \cup \cdots \cup A_k$$

Let B be another event in the sample space given in Figure 4.1. The initial question of interest is how to use the probabilities $P(A_i)$ and $P(B \mid A_i)$ to calculate $P(B)$, the probability of the event B. This can be achieved by noting that

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k)$$

where the events $A_i \cap B$ are mutually exclusive, so that

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k)$$

DRAFT

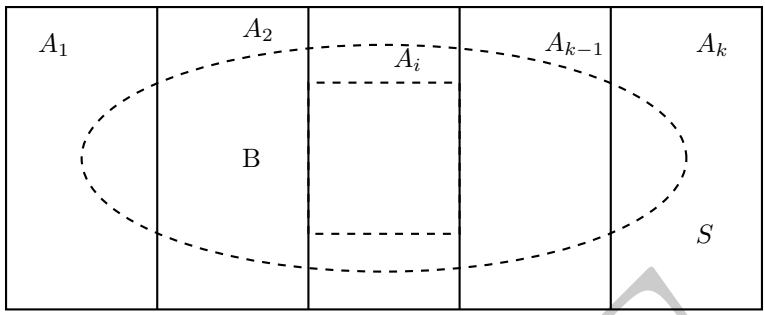


Figure 4.1: A partition $A_1, \ldots, A_k$ and an event B.

Using the definition of conditional probability, this becomes

$$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B \mid A_i)P(A_i) \end{aligned}$$

This result, known as the **Law of Total Probability**, has the interpretation that if it is known that one and only one of a series of events A_i can occur, then the probability of another event B can be obtained as the weighted average of the conditional probabilities $P(B \mid A_i)$, with weights equal to the probabilities $P(A_i)$.

Law of Total Probability: If $A_1, \ldots, A_k$ is a partition of a sample space (i.e., they are mutually exclusive and exhaustive events), then the probability of an event B can be obtained from the probabilities $P(A_i)$ and $P(B \mid A_i)$ using the formula

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)$$

The law of total probability states that if you have a partition of the sample space into mutually exclusive events, the probability of an event can be found by summing the probabilities of the event occurring within each partition, weighted by the probability of each partition.

Problem 4.13. *A car company assembles vehicles at four plants with the following production shares and first-year warranty claim rates:*

Plant (A_i)	***Production Share*** $P(A_i)$	***Claim Rate*** $P(Claim \mid A_i)$
Plant I	*0.20*	*0.05*
Plant II	*0.24*	*0.11*
Plant III	*0.25*	*0.03*
Plant IV	*0.31*	*0.08*

*A **warranty claim** occurs when a buyer reports a defect and the company repairs or replaces the part at no cost within the first year. If a car is selected at random, what is the probability that it will require a warranty claim?*

Solution

To find this, we use the **Law of Total Probability**, combining the chance of getting a car from each plant with the chance that such a car has a problem.

$$\begin{aligned} P(\text{claim}) &= P(\text{claim} \mid \text{Plant I}) \cdot P(\text{Plant I}) + P(\text{claim} \mid \text{Plant II}) \cdot P(\text{Plant II}) \\ &\quad + P(\text{claim} \mid \text{Plant III}) \cdot P(\text{Plant III}) + P(\text{claim} \mid \text{Plant IV}) \cdot P(\text{Plant IV}) \\ &= (0.05 \cdot 0.20) + (0.11 \cdot 0.24) + (0.03 \cdot 0.25) + (0.08 \cdot 0.31) \\ &= 0.01 + 0.0264 + 0.0075 + 0.0248 \\ &= 0.0687 \end{aligned}$$

There is a 6.87% chance that a randomly sold car will need a repair during the first year and the customer will make a warranty claim.

4.7.2 Total Probability with Multiple Conditions

In previous discussions (refer to Section 4.7.1), we introduced the Law of Total Probability, which allows us to compute the probability of an event B by partitioning the sample space based on a single set of mutually exclusive and exhaustive events A_i. However, real-world scenarios often involve more complex situations where the conditions for an event might depend on multiple factors, or where the "simple" partitions A_i themselves are conditional on other events. This section extends the application of the Law of Total Probability to such cases involving multiple conditions or hierarchical dependencies.

Consider a situation where we are interested in the probability of event B, but the likelihood of B occurring depends not just on one set of scenarios A_i, but on a further breakdown within those scenarios, say by another set of events C_m. For example, if A_i represents a product's manufacturing line (e.g., Line 1, Line 2) and B represents a defective product, we might further classify A_i by the shift (e.g., Day Shift, Night Shift), which we can denote as C_1, C_2. So, the probability of a defective product depends on the manufacturing line AND the shift.

To compute $P(B)$ in such a case, we can apply the Law of Total Probability iteratively or by defining a finer partition of the sample space.

Iterative Application

We can apply the Law of Total Probability in stages. Suppose our sample space is partitioned by $A_1, \ldots, A_k$. For each A_i, we might have a further partition

DRAFT

$C_{i1}, C_{i2}, \ldots, C_{im_i}$. The total probability of B can then be written as:

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i)$$

Now, if $P(B \mid A_i)$ itself depends on a sub-condition C_m (meaning, $P(B \mid A_i \cap C_m)$ is what's known), we can use the Law of Total Probability again on $P(B \mid A_i)$. However, it's more precise to directly partition the original sample space into the intersections of the conditions.

Partitioning by Joint Events

A more rigorous approach for multiple conditions is to define a new, finer partition of the sample space using the intersection of all relevant conditions. If we have events $A_1, \ldots, A_k$ that partition the sample space, and another set of events $C_1, \ldots, C_m$ that also partition the sample space, then the intersections $A_i \cap C_j$ form an even finer partition of the sample space. That is, if A_i are mutually exclusive and exhaustive, and C_j are mutually exclusive and exhaustive, then $A_i \cap C_j$ are also mutually exclusive and exhaustive.

In this case, the Law of Total Probability can be expressed as:

$$P(B) = \sum_{i=1}^{k} \sum_{j=1}^{m} P(B \mid A_i \cap C_j)P(A_i \cap C_j)$$

This formula states that to find the probability of B, we sum the conditional probabilities of B given each joint scenario (A_i and C_j simultaneously), weighted by the probability of that joint scenario occurring.

Using the definition of joint probability $P(A_i \cap C_j) = P(C_j \mid A_i)P(A_i)$, we can also write:

$$P(B) = \sum_{i=1}^{k} P(A_i) \sum_{j=1}^{m} P(B \mid A_i \cap C_j)P(C_j \mid A_i)$$

This form is particularly useful when the conditional probabilities of C_j are known given A_i.

Law of Total Probability with Multiple Conditions: If $A_1, \ldots, A_k$ form a partition of the sample space, and $C_1, \ldots, C_m$ also form a partition of the sample space, then for any event B:

$$P(B) = \sum_{i=1}^{k} \sum_{j=1}^{m} P(B \mid A_i \cap C_j)P(A_i \cap C_j)$$

DRAFT

or, if $P(A_i \cap C_j) = P(C_j \mid A_i)P(A_i)$ is more convenient:

$$P(B) = \sum_{i=1}^{k} \sum_{j=1}^{m} P(B \mid A_i \cap C_j)P(C_j \mid A_i)P(A_i)$$

This extended application of the Law of Total Probability is crucial for calculating the evidence $P(B)$, especially when the prior probabilities $P(A_i)$ or likelihoods $P(B \mid A_i)$ are themselves dependent on other factors or intermediate conditions. It allows for a more granular and accurate calculation of probabilities in multi-layered probabilistic models.

Problem 4.14. *In a clinical study, participants are classified by age and smoking status. The probability of being in the young age group (less than 40 years) is* $P(\textit{Young}) = 0.60$*, and the probability of being in the old age group (40 years or older) is* $P(\textit{Old}) = 0.40$*. For the young age group, the conditional probabilities of having high blood pressure (BP) are* $P(\textit{High BP} \mid \textit{Young, Smoker}) = 0.10$ *for smokers and* $P(\textit{High BP} \mid \textit{Young, Non-smoker}) = 0.05$ *for non-smokers. For the old age group, the conditional probabilities of having high BP are* $P(\textit{High BP} \mid \textit{Old, Smoker}) = 0.40$ *for smokers and* $P(\textit{High BP} \mid \textit{Old, Non-smoker}) = 0.25$ *for non-smokers. The probability of being a smoker in the young age group is* $P(\textit{Smoker} \mid \textit{Young}) = 0.30$*, and the probability of being a non-smoker is* $P(\textit{Non-smoker} \mid \textit{Young}) = 0.70$*. In the old age group, the probability of being a smoker is* $P(\textit{Smoker} \mid \textit{Old}) = 0.40$*, and the probability of being a non-smoker is* $P(\textit{Non-smoker} \mid \textit{Old}) = 0.60$*. Calculate the overall probability of having high blood pressure,* $P(\textit{High BP})$*.*

Solution

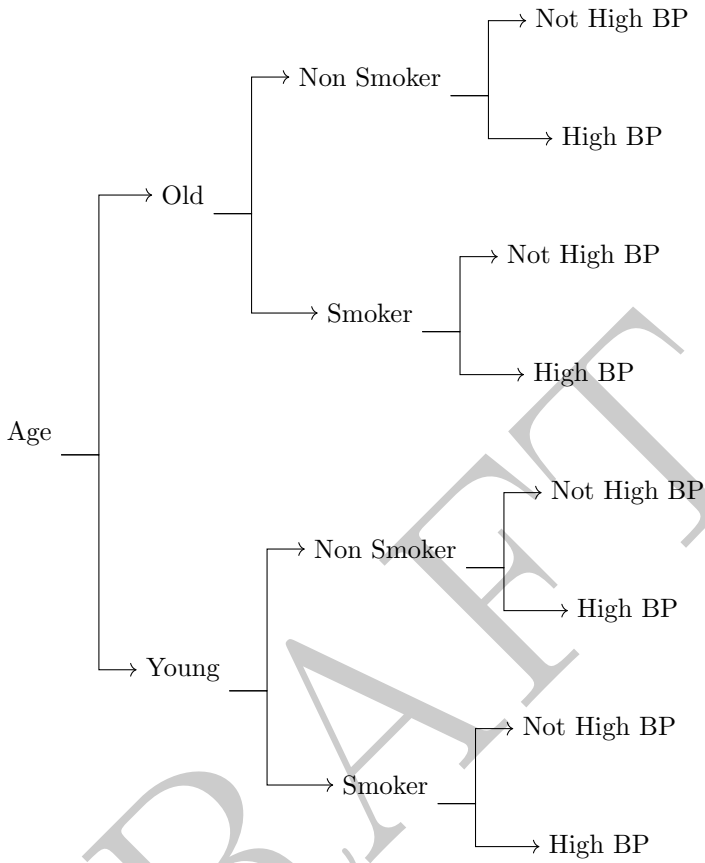
To compute the overall probability of having high blood pressure, $P(\text{High BP})$, we use the Law of Total Probability. The total probability is given by:

$$P(\text{High BP}) = \sum_{i} \sum_{j} P(\text{High BP} \mid A_i, S_j) \cdot P(S_j \mid A_i) \cdot P(A_i)$$

Where

- $P(\text{High BP} \mid A_i, S_j)$ is the conditional probability of having high blood pressure given that the individual belongs to age group A_i and has smoking status S_j,
- $P(A_i)$ is the marginal probability of being in age group A_i,
- $P(S_j \mid A_i)$ is the conditional probability of being a smoker (or non-smoker) given the age group A_i.

DRAFT



Therefore, we have

$$
\begin{aligned}
P(\text{High BP}) &= P(\text{High BP} \mid \text{Young, Smoker}) \cdot P(\text{Smoker} \mid \text{Young}) \cdot P(\text{Young}) \\
&\quad + P(\text{High BP} \mid \text{Young, Non-smoker}) \cdot P(\text{Non-smoker} \mid \text{Young}) \cdot P(\text{Young}) \\
&\quad + P(\text{High BP} \mid \text{Old, Smoker}) \cdot P(\text{Smoker} \mid \text{Old}) \cdot P(\text{Old}) \\
&\quad + P(\text{High BP} \mid \text{Old, Non-smoker}) \cdot P(\text{Non-smoker} \mid \text{Old}) \cdot P(\text{Old}) \\
&= (0.10 \cdot 0.30 \cdot 0.60) + (0.05 \cdot 0.70 \cdot 0.60) + (0.40 \cdot 0.40 \cdot 0.40) \\
&\quad + (0.25 \cdot 0.60 \cdot 0.40) \\
&= 0.018 + 0.021 + 0.064 + 0.060 \\
&= 0.163
\end{aligned}
$$

Thus, the probability of developing high blood pressure is $P(\text{High BP}) = 0.163$ or 16.3%.

Interpretation: The overall probability of having high blood pressure, considering all age groups and smoking statuses, is 16.3%. This result helps understand the prevalence of high blood pressure in the study population, taking

DRAFT

into account various risk factors.

□

4.7.3 Bayes' Theorem

Bayes' Theorem is a cornerstone of probabilistic reasoning and plays a crucial role in many data science applications. At its core, Bayes' Theorem provides a mathematical framework for **updating our beliefs about an event or hypothesis based on new evidence**. This is precisely where the concept of **posterior probability** becomes central.

Bayes' Theorem formally relates the conditional and marginal probabilities of random events. It begins from the fundamental definition of conditional probability:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

By rearranging this, we can solve for $P(A \mid B)$: Let us clarify each term in this fundamental formula:

- $\boldsymbol{P(A \mid B)}$: This is the **Posterior Probability**. It represents the probability of event A being true **after** observing the evidence B. It is our updated belief about A.
- $\boldsymbol{P(B \mid A)}$: This is the **Likelihood**. It represents the probability of observing the evidence B **if** event A is true. It quantifies how well event A explains the observed evidence B.
- $\boldsymbol{P(A)}$: This is the **Prior Probability**. It represents our initial belief or knowledge about the probability of event A **before** any new evidence B is considered.
- $\boldsymbol{P(B)}$: This is the **Marginal Likelihood** or **Evidence**. It represents the total probability of observing event B across all possible scenarios. It acts as a normalizing constant to ensure that all posterior probabilities sum to one.

In cases where the sample space is partitioned by a set of mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_k$, the marginal probability $P(B)$ can be explicitly calculated using the Law of Total Probability:

$$P(B) = \sum_{j=1}^{k} P(B \mid A_j) \cdot P(A_j).$$

Substituting this expanded form of $P(B)$ into the main Bayes' Theorem equation, we get the more comprehensive version, particularly useful when dealing with multiple competing hypotheses A_i:

DRAFT

Bayes' Theorem for Several Mutually Exclusive and Exhaustive Events

If $A_1, A_2, \ldots, A_k$ is a partition of a sample space (i.e., they are mutually exclusive and exhaustive events), then the posterior probability of a specific event A_i given that event B has occurred is given by:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{k} P(B \mid A_j) \cdot P(A_j)}$$

Here, the numerator $P(B \mid A_i) \cdot P(A_i)$ represents the joint probability of B and A_i, and the denominator $\sum_{j=1}^{k} P(B \mid A_j) \cdot P(A_j)$ is the total probability of B occurring.

Bayes' Theorem is particularly useful in scenarios where our initial understanding of probabilities needs to be refined as new data or observations become available. It plays a crucial role in fields such as machine learning (e.g., Naive Bayes classifiers, Bayesian networks), medical diagnosis (calculating the probability of a disease given a test result), and various forms of decision making under uncertainty, by providing a principled way to incorporate new evidence.

Problem 4.15. *Refer to Problem 4.13. A company assembles cars at four plants with the following production shares and warranty claim rates:*

Plant	***Production Share***	***Claim Rate***
Plant I	*20%*	*5%*
Plant II	*24%*	*11%*
Plant III	*25%*	*3%*
Plant IV	*31%*	*8%*

Given that a customer files a warranty claim within the first year, what is the probability that the car came from each plant?

Solution

We already know from Problem 4.13 that the total probability of a claim is:

$$P(\text{Claim}) = 0.0687$$

We now compute the **posterior probabilities** using Bayes' Theorem, i.e., the probability that the car came from a specific plant *given that a claim occurred*:

$$P(\text{Plant I} \mid \text{Claim}) = \frac{P(\text{Plant I}) \cdot P(\text{Claim} \mid \text{Plant I})}{P(\text{Claim})} = \frac{0.20 \times 0.05}{0.0687} = 0.146$$

DRAFT

Applying Bayes' Theorem similarly for the remaining plants:

$$P(\text{Plant II} \mid \text{Claim}) = \frac{P(\text{Plant II}) \cdot P(\text{Claim} \mid \text{Plant II})}{P(\text{Claim})} = \frac{0.24 \times 0.11}{0.0687} = 0.384$$

$$P(\text{Plant III} \mid \text{Claim}) = \frac{P(\text{Plant III}) \cdot P(\text{Claim} \mid \text{Plant III})}{P(\text{Claim})} = \frac{0.25 \times 0.03}{0.0687} = 0.109$$

$$P(\text{Plant IV} \mid \text{Claim}) = \frac{P(\text{Plant IV}) \cdot P(\text{Claim} \mid \text{Plant IV})}{P(\text{Claim})} = \frac{0.31 \times 0.08}{0.0687} = 0.361$$

The results are summarized in the following table:

Plant (A_i)	**Prior** $P(A_i)$	**Likelihood** $P(\text{Claim} \mid A_i)$	**Posterior** $P(A_i \mid \text{Claim})$
Plant I	0.20	0.05	0.146
Plant II	0.24	0.11	0.384
Plant III	0.25	0.03	0.109
Plant IV	0.31	0.08	0.361
Total	1.00		1.000

This means that if a customer makes a warranty claim, the car is most likely to have come from **Plant II**, which has the highest claim rate (11%). Before the claim, Plant II accounted for only 24% of production. But after observing the claim, there is a 38.4% chance it came from Plant II. In contrast, Plant III had the lowest claim rate (3%), so the probability that a claimed car came from there drops from 25% to just 10.9%. This is the power of Bayes' Theorem, it allows us to *update* our beliefs about the source of an event after observing new evidence. □

Problem 4.16. *Suppose it is known that 1% of the population suffers from a particular disease. A blood test has a 97% chance of identifying the disease for diseased individuals, but also has a 6% chance of falsely indicating that a healthy person has the disease.*

(a) What is the probability that a person will have a positive blood test?

(b) If your blood test is positive, what is the chance that you have the disease?

(c) If your blood test is negative, what is the chance that you do not have the disease?

DRAFT

Solution

Let's define the events

- D^+ = person has the disease, so $P(D^+) = 0.01$, $P(D^-) = 0.99$
- T^+ = blood test is positive
- T^- = blood test is negative

Given information that

$$P(T^+ \mid D^+) = 0.97, \qquad P(T^+ \mid D^-) = 0.06$$

Part (a): $P(T^+)$

Using the **Law of Total Probability**, we have

$$\begin{aligned} P(T^+) &= P(T^+ \mid D^+)\, P(D^+) + P(T^+ \mid D^-)\, P(D^-) \\ &= (0.97)(0.01) + (0.06)(0.99) = 0.0691 \end{aligned}$$

Part (b): $P(D^+ \mid T^+)$

Using **Bayes' Theorem**, we have

$$P(D^+ \mid T^+) = \frac{P(T^+ \mid D^+)\, P(D^+)}{P(T^+)} = \frac{(0.97)(0.01)}{0.0691} \approx 0.1404$$

There is only about a **14% chance** that you actually have the disease given a positive test result. This may seem surprisingly low, but it is explained by the fact that the disease is rare, most positive tests come from the large pool of healthy individuals.

Part (c): $P(D^- \mid T^-)$

Using the Law of Total Probability:

$$\begin{aligned} P(T^-) &= P(T^- | D^+)P(D^+) + P(T^- | D^-)P(D^-) \\ &= 0.03 \times 0.01 + 0.94 \times 0.99 \\ &= 0.0003 + 0.9306 \\ &= 0.9309 \end{aligned}$$

(Alternatively, $P(T^-) = 1 - P(T^+) = 1 - 0.0691 = 0.9309$, which matches.) Also, $P(T^- \mid D^-) = 1 - 0.06 = 0.94$. By Bayes' Theorem, we have

$$P(D^- \mid T^-) = \frac{P(T^- \mid D^-)\, P(D^-)}{P(T^-)} = \frac{(0.94)(0.99)}{0.9309} \approx 0.9997$$

If your blood test is negative, there is approximately a **99.97% chance** that you do not have the disease. □

DRAFT

4.8 Concluding Remarks

In this chapter, we covered the essential principles of probability that form the backbone of data science. We discussed experiments, sample spaces, joint and marginal probabilities, and conditional probabilities, providing a solid foundation for analyzing uncertainty and making data-driven decisions. We also explored various methods for assigning probabilities, including classical, empirical, and subjective approaches. The insights gained from understanding joint probabilities, marginal probabilities, and Bayes' Theorem will be invaluable for refining models and interpreting data.

As we move forward, the next chapter will delve into random variables and their properties. Random variables are crucial for quantifying and modeling uncertainty in a more structured way. We will explore different types of random variables, their distributions, and key properties, further building on the probability concepts introduced here. Mastering these topics will enhance your ability to handle complex data challenges and apply statistical techniques effectively. Understanding random variables is essential for advanced data analysis and developing predictive models.

4.9 Chapter Exercises

1. Consider an experiment where a fair six-sided die is rolled. Define the following events:

   - A: The event that the outcome is an even number.
   - B: The event that the outcome is greater than 4.

   Calculate the following probabilities:

   (a) $P(A)$
   (b) $P(B)$
   (c) $P(A \cap B)$
   (d) $P(A \cup B)$
   (e) $P(A^c)$

2. In a bag of 10 balls, 4 are red and 6 are blue. Two balls are drawn at random without replacement. Define the following events:

   - A: Drawing a red ball on the first draw.
   - B: Drawing a red ball on the second draw.

   Calculate the following probabilities:

   (a) $P(A)$

DRAFT

(b) $P(B \mid A)$
(c) $P(A \cap B)$
(d) $P(A \cup B)$

3. You are given a deck of 52 playing cards. Define the following events:

   - A: Drawing a card that is a heart.
   - B: Drawing a card that is a queen.

   Calculate the following probabilities:

   (a) $P(A)$
   (b) $P(B)$
   (c) $P(A \cap B)$
   (d) $P(A \cup B)$
   (e) $P(A^c)$

4. Suppose that the probability of a young person liking Facebook is 0.7, the probability of liking YouTube is 0.6, and the probability of liking both platforms is 0.5. Using the relevant probability theorems, determine the following:

   (a) What is the probability that a young person likes **exactly one** of the two social media platforms?
   (b) What is the probability that a young person likes **at least one** of the two platforms?
   (c) What is the probability that a young person likes **only Facebook** and **not YouTube**?

5. A quality control team in a small factory inspects a batch of 60 parts. It was observed that 10 parts were defective in appearance, 12 parts had functional defects, and 4 parts were both defective in appearance and function. If a part is selected randomly, what is the probability that it is defective in appearance or has a functional defect?

6. A survey finds that 60% of people prefer coffee, 50% prefer tea, and 30% prefer both coffee and tea. What is the probability that a randomly chosen person prefers **at least one** of the two drinks?

7. In a class of 30 students, 18 like mathematics, 12 like science, and 8 like both. If a student is chosen at random, calculate:

   (a) The probability that the student likes science.
   (b) The probability that the student likes mathematics given they like science.

DRAFT

(c) The probability that the student likes either mathematics or science.

8. Consider the employment status of male and female employees in a technology company. The company employs 1500 individuals, of whom 1050 are men and 450 are women. Over the past year, 375 employees were promoted. After analyzing the promotion records, a committee of female employees raised concerns about potential gender bias, noting that 315 male employees had received promotions, while only 60 female employees were promoted.

   (i) Construct a joint probability table based on these data. Calculate the marginal probabilities. If a male employee is selected randomly, what is the probability that he was promoted?

   (ii) If a female employee is selected randomly, what is the probability that she was not promoted? Also, if a randomly selected employee was promoted, what is the probability that the employee is male?

   (iii) If a randomly selected employee was not promoted, what is the probability that the employee is female?

9. A team has a probability of 0.60 of winning a match. The probability that the team goes the whole match without a red card is 0.80. If they avoid a red card, the probability that they win is 0.70.

   (a) Find the probability that the team wins and has no red card.

   (b) Find the probability that the team gets a red card but still wins the match.

10. Among those people who are infected with a certain virus, 32% have strain A, 59% have strain B, and the remaining 9% have strain C. Furthermore, 21% of people infected with strain A of the virus exhibit symptoms, 16% of people infected with strain B of the virus exhibit symptoms, and 63% of people infected with strain C of the virus exhibit symptoms.

   (a) If a person has the virus and exhibits symptoms of it, what is the probability that they have strain C?

   (b) If a person has the virus but doesn't exhibit any symptoms of it, what is the probability that they have strain A?

   (c) What is the probability that a person who has the virus does not exhibit any symptoms of it?

11. Suppose 1% of the population in a country are HIV-positive. A test procedure for diagnosing whether a person is HIV-positive indicates with probability 0.98 that the person is HIV-positive if indeed he/she is HIV-positive, and with probability 0.96 that this person is not HID-positive if he/she is not HIV-positive.

DRAFT

(i) What is the probability that a test person is HIV-positive if the test indicates that?

(ii) What is the probability that a person is HIV-negative if the test procedure indicates this?

12. A company has a fleet of five delivery trucks. The probability that Truck 1 is available on a given day is 0.95, Truck 2 is available with a probability of 0.90, Truck 3 with a probability of 0.85, Truck 4 with a probability of 0.88, and Truck 5 with a probability of 0.92. Assume the availability of each truck is independent of the others.

    (a) The company's main route requires all five trucks to be available. What's the probability that the main route can be serviced on a given day?

    (b) A smaller local route can be serviced if at least one truck is available. What's the probability that the local route can be serviced?

    (c) A contract with a major client requires at least three trucks to be available. What's the probability that the company can fulfill this contract?

13. A water-supply system consists of four pumps labelled A, B, C, and D. The system delivers water if pump A works *and* pump B works *and* at least one of pumps C or D works (i.e. C and D are in parallel, and this parallel pair is in series with A and B). The reliabilities of the pumps are: $P(A) = 0.92$, $P(B) = 0.88$, $P(C) = 0.76$, $P(D) = 0.82$. Assume the four pumps operate independently. Calculate:

    (i). The probability that the entire system delivers water.

    (ii). The probability that pump C fails, given that the entire system delivers water.

    (iii). The probability that pump D fails, given that the entire system delivers water.

14. In a clinical study, researchers are interested in the probability of a patient developing a particular health condition based on the type of treatment received. There are three types of treatments: A, B, and C. The probabilities of receiving each treatment are as follows:

    - Treatment A: 30% ($P(A) = 0.30$)
    - Treatment B: 50% ($P(B) = 0.50$)
    - Treatment C: 20% ($P(C) = 0.20$)

    The probability of developing the health condition given the type of treatment is known to be:

DRAFT

- $P(\text{Condition} \mid A) = 0.10$
- $P(\text{Condition} \mid B) = 0.25$
- $P(\text{Condition} \mid C) = 0.15$

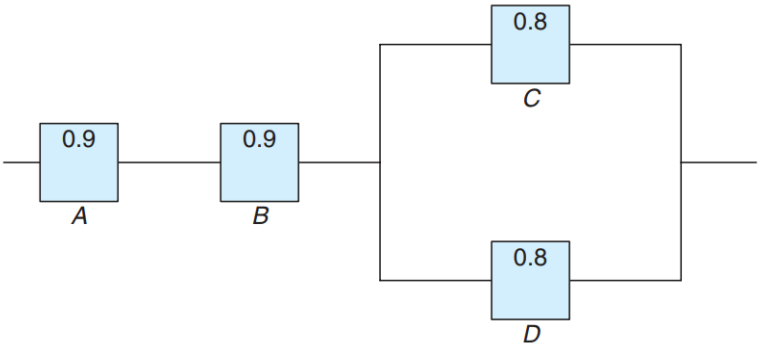
Find the overall probability of a patient developing the health condition, denoted as $P(\text{Condition})$.

Using Bayes' theorem, calculate the following probabilities:

(a) $P(A \mid \text{Condition})$
(b) $P(B \mid \text{Condition})$
(c) $P(C \mid \text{Condition})$.

15. Suppose that somebody secretly rolls two fair six-sided dice, and what is the probability that the face-up value of the first one is 3, given the information that their sum is no greater than 5?

16. An electrical system consists of four components as illustrated in the following figure.



The system works if components A and B work and either of the components C or D works. The reliability (probability of working) of each component is also shown in the above figure. Assume that the four components work independently. Calculate the following probabilities:

(a) Probability that the entire system works.
(b) Probability that component C does not work, given the entire system works.
(c) Probability that component D does not work, given the entire system works.

DRAFT

17. An agricultural research establishment grows vegetables and grades each one as either good or bad for its taste, good or bad for its size, and good or bad for its appearance. Overall 78% of the vegetables have a good taste. However, only 69% of the vegetables have both a good taste and a good size. Also, 5% of the vegetable have both a good taste and a good appearance, but a bad size. Finally, 84% of the vegetables have either a good size or a good appearance.

    (a). If a vegetable has a good taste, what is the probability that it also has a good size?

    (b). If a vegetable has a bad size and a bad appearance, what is the probability that it has a good taste?

    .

18. A company produces electronic components, and it has two types of machines, A and B, that manufacture these components. Machine A produces 60% of the components, while Machine B produces 40%. Historical data shows that 2% of the components produced by Machine A are defective, while 5% of the components produced by Machine B are defective.

    A component is selected at random and found to be defective. What is the probability that this defective component was produced by Machine A?

19. A certain rare disease affects 2% of a population. A diagnostic test for this disease has the following characteristics:

    - It correctly identifies the disease (true positive) 90% of the time for those who have it.
    - It incorrectly indicates the disease (false positive) 5% of the time for those who do not have it.

    Answer the following questions.

    (a) What is the probability that a person will receive a positive test result?

    (b) If a person tests positive, what is the probability that they actually have the disease?

    (c) If a person tests negative, what is the probability that they do not have the disease?

20. As a mining company evaluates the likelihood of discovering a gold deposit in a specific region, they have gathered data on the probabilities associated with geological features. Given that the probability of finding a gold deposit is $P(G) = 0.3$, the likelihood of observing specific geological features if a deposit is present is $P(E|G) = 0.8$, and the chance of

DRAFT

observing those features if no deposit exists is $P(E|G^c) = 0.1$, answer the following:

(i) Calculate the probability of observing the geological features in this area.

(ii) What is the probability that there is indeed a gold deposit given the observed geological features?

(iii) How would you interpret these results in terms of their implications for the mining company's decision-making process regarding further exploration in this region?

21. A factory produces 80% of products with Machine A and 20% with Machine B. If 2% of A's products and 5% of B's products are defective, what is the probability that a defective product came from Machine A?

22. Suppose you are on a game show with three doors: one has a car, the other two have goats. You choose Door 1. The host, who knows what's behind the doors, opens Door 3 to show a goat and asks if you want to switch to Door 2.

    (a) What's the chance of winning the car if you switch to Door 2?

    (b) What's the chance of winning the car if you stay with Door 1?

DRAFT

Chapter 5

Random Variable and Its Properties

DRAFT

5.1 Introduction

In the realm of data science, understanding and manipulating uncertainty is a fundamental skill. At the core of this capability lies the concept of a random variable. A random variable is a quantitative variable whose values are determined by the outcome of a random phenomenon. It serves as a bridge connecting the abstract world of probability theory to the concrete domain of data analysis.

Random variables can be classified into two main types: discrete and continuous. Discrete random variables take on a countable number of distinct values, often representing things like the number of occurrences of an event. Continuous random variables, on the other hand, can take on an infinite number of possible values within a given range, making them essential for representing measurements and other quantities that vary smoothly.

This chapter delves into the foundational aspects of random variables, exploring their properties and the critical role they play in statistical modeling and data analysis. We will discuss probability distributions, expected values, variances, and other By the end of this chapter, readers will gain a robust understanding of how random variables function and how they can be applied to solve real-world problems in data science.

5.2 Random Variable

A **random variable** is a mathematical concept used in probability theory and statistics, representing a numerical outcome whose possible values depend on

a random experiment. It serves as a fundamental tool for defining probability distributions and calculating probabilities associated with events arising from uncertain or stochastic processes. In data science, random variables are fundamental because they allow us to model and reason about uncertainty and variability in data.

A random variable is a function that assigns a real number to each outcome in a sample space of a random experiment. Formally, a random variable X is defined as a function:

$$X : S \to \mathbb{R}$$

where S is the sample space of the experiment, and $\mathbb{R}$ is the set of real numbers.

Random Variable: A variable that assigns numerical values to outcomes of a random experiment, with each value having a specific probability.

Example: Testing Electronic Components

Consider a **random experiment** where three electronic components are tested for defects. The sample space, giving a detailed description of each possible outcome, can be written as follows:

$$S = \{\text{NNN}, \text{NDN}, \text{NND}, \text{DNN}, \text{NDD}, \text{DND}, \text{DDN}, \text{DDD}\}$$

where,

- N stands for a non-defective component.
- D stands for a defective component.

Let X be the random variable representing the **number of defective components** in the sample. The possible values of X are denoted by x, and their corresponding outcomes are listed in Table 5.1. The random variable X can take on the following values:

- $x = 0$: Corresponds to the outcome NNN (No defective components)
- $x = 1$: Corresponds to the outcomes {NDN, NND, DNN} (One defective component)
- $x = 2$: Corresponds to the outcomes {NDD, DND, DDN} (Two defective components)
- $x = 3$: Corresponds to the outcome DDD (Three defective components).

DRAFT

Table 5.1: Possible Outcomes When Testing Three Electronic Components

Outcome	NNN	NDN	NND	DNN	NDD	DND	DDN	DDD
x	0	1	1	1	2	2	2	3

In this example, X is a **discrete random variable** because it can take on a countable number of distinct values. Each value of X corresponds to the number of defective components in the tested sample. There are two main types of random variables: discrete and continuous.

5.3 Discrete Random Variables

A **discrete random variable** can take on a countable number of possible values. Here are some examples of discrete random variables:

1. **Number of Heads in a Series of Coin Tosses**: When flipping a fair coin multiple times, the number of heads observed in the series is a discrete random variable. For example, if you flip a coin 10 times, the number of heads (0 to 10) is a discrete outcome.

2. **Number of Defective Items in a Batch**: In quality control, the number of defective items in a batch of products is a discrete random variable. For instance, if a factory produces 100 items in a day, the number of defective items could be any integer from 0 to 100.

3. **Number of Customers in a Queue**: The number of customers waiting in line at a service center or a bank is a discrete random variable. At any given time, this number could be 0, 1, 2, and so on.

4. **Roll of a Die**: When rolling a standard six-sided die, the outcome is a discrete random variable with possible values of 1, 2, 3, 4, 5, or 6.

5. **Number of Emails Received in a Day**: The number of emails a person receives in a day is a discrete random variable. It can take on any non-negative integer value (0, 1, 2, ...).

6. **Number of Accidents at an Intersection**: The number of traffic accidents occurring at a particular intersection in a month is a discrete random variable. This count could be 0, 1, 2, and so on.

7. **Number of Children in a Family**: The number of children in a family is a discrete random variable, with possible values of 0, 1, 2, and so forth.

8. **Number of Sales Transactions in a Day**: The number of sales transactions processed by a retail store in a single day is a discrete random variable, representing the count of individual sales.

DRAFT

These examples illustrate various contexts in which discrete random variables are used to model and analyze real-world phenomena.

5.3.1 Probability Mass Function (PMF)

The probability distribution of a random variable describes how probabilities are distributed over the values of the random variable. For a discrete random variable, the probability distribution is described by the **probability mass function (PMF)**, which gives the probability that the random variable takes on a specific value.

Probability Mass Function: A PMF of a discrete random variable X is a function $p(x)$ that gives the probability that X takes the value x. It satisfies:

(i). **Non-negativity**: For every possible value x that X can take, the probability $p(x)$ is non-negative:

$$p(x) \geq 0$$

(ii). **Normalization**: The sum of the probabilities for all possible values of X equals 1:

$$\sum_{x \in \text{Range of } X} p(x) = 1$$

(iii). **Probability Assignment**: For any specific value x, $p(x)$ gives the probability that the random variable X takes the value x:

$$p(x) = P(X = x)$$

Example: Testing Electronic Components

Consider the example of **Testing Electronic Components** described in the previous section, where X is the random variable representing the **number of defective components** in the three tested electronic components. The probability mass function for X is shown below, and the graphical representation is presented in Figure 5.1.

Table 5.2: Probability mass function

Outcome	x	$P(X = x)$
{NNN}	0	$\frac{1}{8}$
{NDN, NND, DNN}	1	$\frac{3}{8}$
{NDD, DND, DDN}	2	$\frac{3}{8}$
{DDD}	3	$\frac{1}{8}$

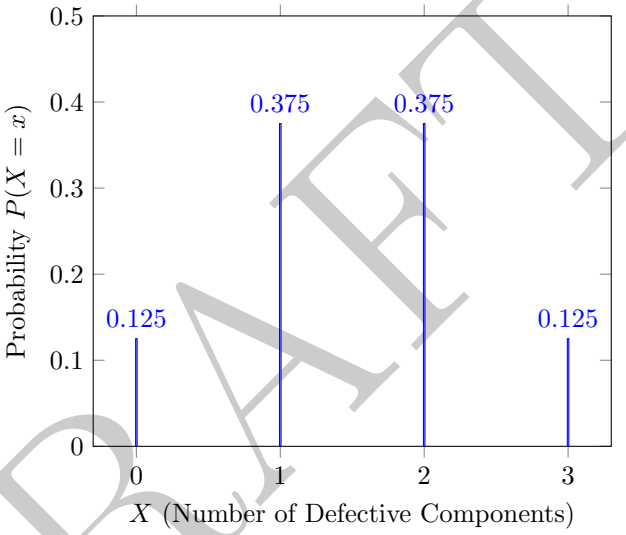


Figure 5.1: Probability Mass Function of X

5.3.2 Cumulative Distribution Function (CDF)

The **cumulative distribution function** (CDF) of a random variable X is a function that gives the probability that X will take a value less than or equal to x. For both discrete and continuous random variables,

$$F(x) = P(X \leq x).$$

The CDF is a non-decreasing function that ranges from 0 to 1.

For the discrete random variable X, the cumulative distribution function can then be calculated from the expression:

$$F(x) = \sum_{y:y \leq x} P(X = y).$$

DRAFT

Example: Testing Electronic Components

Consider the example of **Testing Electronic Components** described in the previous section, where X is the random variable representing the **number of defective components** in the three tested electronic components. The cumulative distribution function for X is shown below, and the graphical representation is presented in Figure 5.2.

Table 5.3: The CDF for the Number of Defective Components

x	$P(X = x)$	$F(x)$
0	$\frac{1}{8}$	0.125
1	$\frac{3}{8}$	0.500
2	$\frac{3}{8}$	0.875
3	$\frac{1}{8}$	1

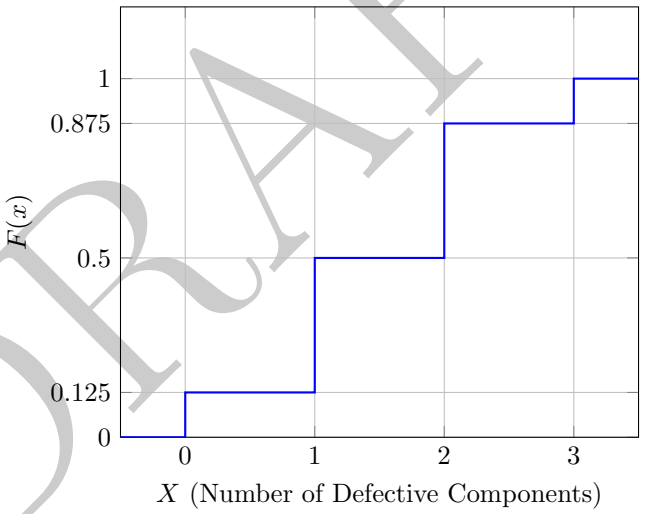


Figure 5.2: Cumulative Distribution Function of X

DRAFT

5.3.3 Properties of the Cumulative Distribution Function

The CDF of a random variable has several important properties:

1. **Non-decreasing**: The CDF $F(x)$ is a non-decreasing function. This means that if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$. The probability that the random variable takes a value less than or equal to x does not decrease as x increases.

2. **Limits**:
   - $\lim_{x \to -\infty} F(x) = 0$; that is the minimum value of the CDF is 0.
   - $\lim_{x \to +\infty} F(x) = 1$; that is the maximum value of the CDF is 1.
3. **Right-Continuous**: The CDF $F(x)$ is right-continuous. This means that for any value x, the limit of $F(x)$ as t approaches x from the right ($t \to x^+$) is equal to $F(x)$. Mathematically, this can be written as $\lim_{t \to x^+} F(t) = F(x)$.
4. **Range**: The CDF $F(x)$ takes values in the interval $[0, 1]$. For any real number x, $0 \leq F(x) \leq 1$. This reflects the fact that probabilities range from 0 to 1.
5. **Step Function**: For discrete random variables, the CDF $F(x)$ is a **step function**, with the value increasing at each point where the random variable takes a value.

Problem 5.1. *An office has four copying machines, and the random variable X measures how many of them are in use at a particular moment in time. Suppose that $P(X = 0) = 0.08$, $P(X = 1) = 0.11$, $P(X = 2) = 0.27$, and $P(X = 3) = 0.33$.*

(a) What is $P(X = 4)$?

(b) Draw a line graph of the probability mass function.

(c) Construct and plot the cumulative distribution function.

Solution

(a) Since the sum of all probabilities must be 1, we have:

$$\begin{aligned} P(X = 4) &= 1 - (P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)) \\ &= 1 - (0.08 + 0.11 + 0.27 + 0.33) = 1 - 0.79 \\ &= 0.21 \end{aligned}$$

(b) The graphical presentation of the probability mass function is the following:

DRAFT

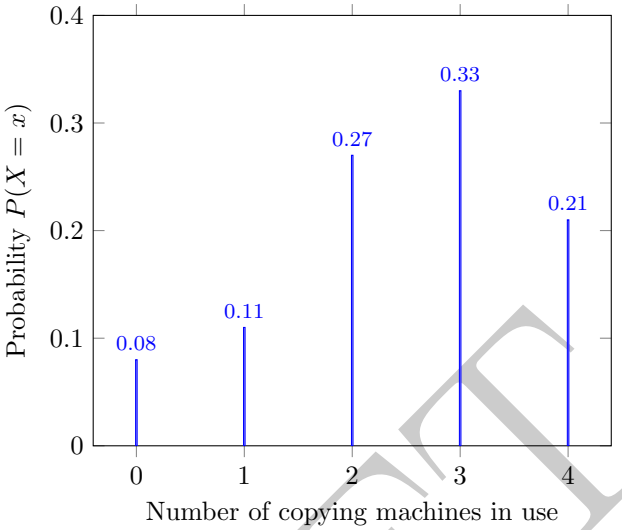


Figure 5.3: Probability Mass Function

(c) We konw, the cumulative distribution function $F(x)$ is defined as:

$$F(x) = P(X \leq x)$$

The cumulative distribution function $F(x)$ with probability mass function is provided in Table 5.4.

x	0	1	2	3	4
$p(x)$	0.08	0.11	0.27	0.33	0.21
$F(x)$	0.08	0.19	0.46	0.79	1.00

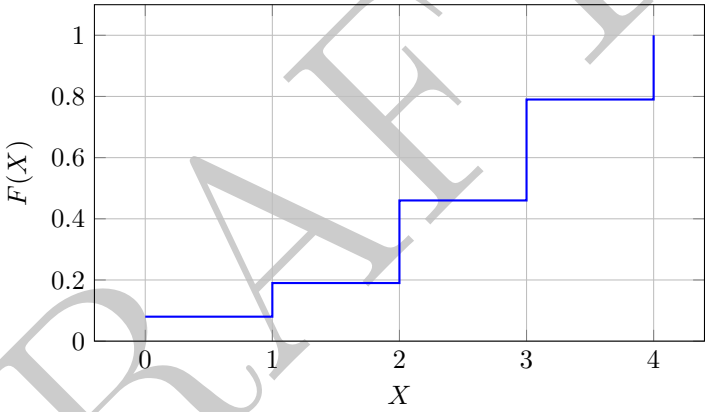
Table 5.4: Cumulative Distribution Function of X

DRAFT

where,

$$
\begin{aligned}
F(0) &= P(X = 0) = 0.08 \\
F(1) &= P(X \leq 1) = P(X = 0) + P(X = 1) = 0.08 + 0.11 = 0.19 \\
F(2) &= P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \\
&= 0.08 + 0.11 + 0.27 = 0.46 \\
F(3) &= P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\
&= 0.08 + 0.11 + 0.27 + 0.33 = 0.79 \\
F(4) &= P(X \leq 4) \\
&= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\
&= 0.08 + 0.11 + 0.27 + 0.33 + 0.21 = 1.00
\end{aligned}
$$

The graphical presentation of $F(x)$ is the following:



Problem 5.2. *Let the number of phone calls received by a switchboard during a 5-minute interval be a random variable X with probability function*

$$p(x) = \frac{e^{-2}2^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots$$

(a) Determine the probability that x equals 0, 1, 2, 3, 4, 5, and 6.

(b) Graph the probability mass function for these values of x.

(c) Determine the cumulative distribution function for these values of X.

Solution

Let X be a discrete random variable representing the number of phone calls received by a switchboard during a 5-minute interval. The probability function is given by

$$p(x) = \frac{e^{-2}2^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots$$

DRAFT

(a) Probabilities

The probabilities for $X = 0, 1, 2, 3, 4, 5, 6$ are:

$$P(X=0) = \frac{e^{-2}2^0}{0!} = e^{-2} \approx 0.1353$$
$$P(X=1) = \frac{e^{-2}2^1}{1!} = 2e^{-2} \approx 0.2707$$
$$P(X=2) = \frac{e^{-2}2^2}{2!} = 2^2e^{-2}/2 \approx 0.2707$$
$$P(X=3) = \frac{e^{-2}2^3}{3!} = 2^3e^{-2}/6 \approx 0.1804$$
$$P(X=4) = \frac{e^{-2}2^4}{4!} = 2^4e^{-2}/24 \approx 0.0902$$
$$P(X=5) = \frac{e^{-2}2^5}{5!} = 2^5e^{-2}/120 \approx 0.0361$$
$$P(X=6) = \frac{e^{-2}2^6}{6!} = 2^6e^{-2}/720 \approx 0.0120$$

(b) Graph of the Probability Mass Function

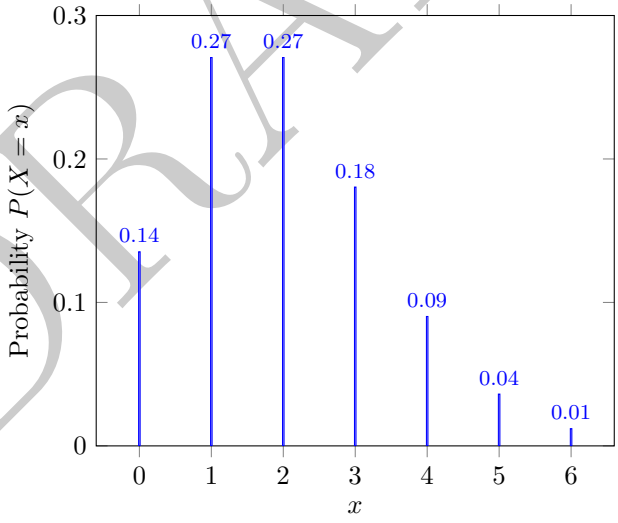


Figure 5.4: Probability Mass Function of X

DRAFT

(c) Cumulative Distribution Function

The **cumulative distribution function** of X is defined as:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \frac{e^{-2} \cdot 2^k}{k!}$$

We now compute both the PMF and the CDF for $x = 0, 1, 2, 3, 4, 5, 6$. The values of $e^{-2} \approx 0.1353$ are used for numerical approximation.

Value of x	0	1	2	3	4	5	6
$P(x)$	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120
$F(x)$	0.1353	0.4060	0.6767	0.8571	0.9473	0.9834	0.9954

5.3.4 Exercise

1. An office has four copying machines, and the random variable X denotes how many of them are in use in a particular time. Suppose the probability mass function X is given below:

x	0	1	2	3	4
$\Pr(X = x)$	k	0.02	0.05	0.4	$(k + 0.3)$

   (a) What is the value of k and draw the line graph of the probability mass function $\Pr(X = x)$.

   (b) Find the Value of $\Pr(X \leq 2)$.

   (c) Find the probability that at least two copying machines are in use.

   (d) Find the cumulative Function $F(x)$ and draw the $F(x)$.

2. An office has five printers and the random variable Y measures how many of them are currently being used. Suppose that $P(Y = 0) = 0.05$, $P(Y = 1) = 0.10$, $P(Y = 2) = 0.20$, $P(Y = 3) = 0.30$, and $P(Y = 4) = 0.25$.

   (a) What is $P(Y = 5)$?

   (b) Draw a line graph of the probability mass function.

   (c) Construct and plot the cumulative distribution function.

3. A hospital has six emergency rooms and the random variable Z measures how many of them are occupied at a given time. Suppose that $P(Z = 0) = 0.04$, $P(Z = 1) = 0.10$, $P(Z = 2) = 0.20$, $P(Z = 3) = 0.25$, $P(Z = 4) = 0.20$, and $P(Z = 5) = 0.15$.

DRAFT

(a) What is $P(Z = 6)$?

(b) Draw a line graph of the probability mass function.

(c) Construct and plot the cumulative distribution function.

4. A hospital has three emergency rooms, and the random variable W denotes how many of them are occupied at a particular time. Suppose the probability mass function W is given below:

w	0	1	2	3
$\Pr(W = w)$	q	0.15	0.25	$(q + 0.05)$

(a) What is the value of q and draw the line graph of the probability mass function $\Pr(W = w)$.

(b) Find the value of $\Pr(W \leq 2)$. Also, find the probability that at least one emergency room is occupied.

(c) Find the cumulative distribution function $F(w)$ and draw the graph of $F(w)$.

5. A clinic has three doctors and the random variable W measures how many of them are available at a particular moment in time. Suppose that $P(W = 0) = 0.15$, $P(W = 1) = 0.20$, and $P(W = 2) = 0.30$.

(a) What is $P(W = 3)$?

(b) Draw a line graph of the probability mass function.

(c) Construct and plot the cumulative distribution function.

6. A warehouse has seven forklifts and the random variable V measures how many of them are currently in operation. Suppose that $P(V = 0) = 0.02$, $P(V = 1) = 0.08$, $P(V = 2) = 0.18$, $P(V = 3) = 0.25$, $P(V = 4) = 0.20$, and $P(V = 5) = 0.15$.

(a) What is $P(V = 6)$?

(b) Draw a line graph of the probability mass function.

(c) Construct and plot the cumulative distribution function.

7. A manufacturing plant has four assembly lines and the random variable U measures how many of them are operating at a given time. Suppose that $P(U = 0) = 0.10$, $P(U = 1) = 0.20$, and $P(U = 2) = 0.35$.

(a) What is $P(U = 3)$?

(b) Draw a line graph of the probability mass function.

(c) Construct and plot the cumulative distribution function.

DRAFT

5.4 Continuous Random Variables

A **continuous random variable** can take on an uncountable number of possible values. Here are some examples of continuous random variables:

1. **Height of Individuals**: The height of a person is a continuous random variable because it can take any value within a given range. For example, the height could be 170.2 cm, 175.5 cm, etc.

2. **Time Taken to Complete a Task**: The time required to finish a task, such as running a marathon, is a continuous random variable. It can be measured in hours, minutes, seconds, and fractions of a second.

3. **Temperature**: The temperature at a specific location and time is a continuous random variable. It can take any value within the possible range of temperatures, such as 23.45°C, 37.8°C, etc.

4. **Weight of an Object**: The weight of an object is a continuous random variable. For example, a bag of flour might weigh 1.25 kg, 1.30 kg, etc.

5. **Amount of Rainfall**: The amount of rainfall in a day is a continuous random variable. It can be measured in millimeters or inches, and it can take any value within a range.

6. **Price of a Stock**: The price of a stock at any given moment is a continuous random variable. It can vary continuously and take on any value within the range of possible stock prices.

7. **Age of an Individual**: The age of a person can be considered a continuous random variable if measured precisely. For instance, someone could be 25.3 years old, 45.7 years old, etc.

8. **Voltage in an Electrical Circuit**: The voltage at a point in an electrical circuit is a continuous random variable. It can take any value within the possible voltage range.

These examples illustrate various contexts in which continuous random variables are used to model and analyze real-world phenomena.

5.4.1 Probability Density Function (PDF)

For a continuous random variable X, the probability distribution is described by the **probability density function** (PDF), denoted as $f(x)$. This function specifies the probability density at each point in the random variable's range. The PDF $f(x)$ has the following properties:

1. **Non-negativity**: $f(x) \geq 0$ for all x.

DRAFT

2. **Normalization**: The total area under the PDF curve over the entire range of X is equal to 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Note that the PDF $f(x)$ does not provide the probability of X taking any specific value (which is always zero for continuous random variables). Instead, it indicates the density of probability at each point. To find the probability that X falls within a specific interval $[a, b]$ given in Figure 5.5, you integrate the PDF over that interval:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

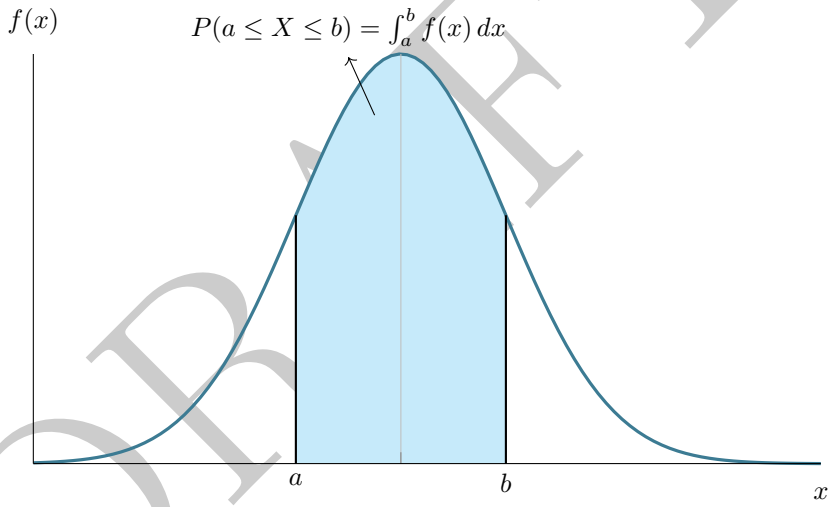


Figure 5.5: The area under the probability density function $f(x)$ between a and b.

Sometimes, the density of X is denoted by $f_X(x)$ to explicitly indicate that the function f corresponds to the random variable X.

Probability Density Function: The probability density function (PDF) of a continuous random variable X with support S is an integrable function $f(x)$ satisfying the following conditions:

(i). $f(x)$ is positive everywhere in the support S, that is, $f(x) \geq 0$ for all $x \in S$. The area under the curve $f(x)$ in the support S is 1,

DRAFT

that is:

$$\int_S f(x)\,dx = 1.$$

(ii). If $f(x)$ is the PDF of X, then the probability that X belongs to an interval $[a, b]$ is given by the integral of $f(x)$ over that interval, that is:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

It is useful to notice that the probability that a continuous random variable X takes any specific value a is always 0! Technically, this can be seen by noting that

$$P(X = a) = \int_a^a f(x)\,dx = 0.$$

Problem 5.3 (Metal Cylinder Production)**.** *Suppose the diameter X of a metal cylinder has the following probability density function (PDF):*

$$f(x) = \begin{cases} 1.5 - 6(x - 50.0)^2 & \text{for } 49.5 \leq x \leq 50.5 \\ 0 & \text{otherwise.} \end{cases}$$

(i). Prove that $f(x)$ is a valid probability density function.

(ii). Find the probability that the diameter of the metal cylinder lies between 49.8 mm and 50.1 mm, i.e., calculate $P(49.8 \leq X \leq 50.1)$.

Solution

(i). To determine if $f(x) = 1.5-6(x-50.0)^2$ for $49.5 \leq x \leq 50.5$ and $f(x) = 0$ elsewhere is a valid probability PDF, we need to check two conditions:

1. **Non-negativity**: $f(x) \geq 0$ for all x.
2. **Normalization**: The total integral of $f(x)$ over all possible values must equal 1.

Non-negativity Check

We need to ensure that $f(x) \geq 0$ for $49.5 \leq x \leq 50.5$:

$$f(x) = 1.5 - 6(x - 50.0)^2$$

For $49.5 \leq x \leq 50.5$, let's calculate the minimum value of the quadratic function: $(x - 50.0)^2$ is minimized at $x = 50.0$, and $(x - 50.0)^2$ ranges from 0 to $(0.5)^2 = 0.25$.

$$f(x) = 1.5 - 6(x - 50.0)^2 \geq 1.5 - 6 \cdot 0.25 = 1.5 - 1.5 = 0.$$

Thus, $f(x) \geq 0$ for all x in the given interval.

DRAFT

Normalization Check

We need to integrate $f(x)$ over the interval $49.5 \leq x \leq 50.5$ and check if the integral equals 1:

$$\begin{aligned}\int_{49.5}^{50.5} f(x)\,dx &= \int_{49.5}^{50.5} \left(1.5 - 6(x-50.0)^2\right) dx \\ &= \int_{49.5}^{50.5} 1.5\,dx - \int_{49.5}^{50.5} 6(x-50.0)^2\,dx \\ &= 1.5 \times (50.5 - 49.5) - \int_{49.5}^{50.5} 6(x-50.0)^2\,dx \\ &= 1.5 - \int_{49.5}^{50.5} 6(x-50.0)^2\,dx. \qquad (5.1)\end{aligned}$$

(5.2)

The integral

$$\int_{49.5}^{50.5} 6(x-50.0)^2\,dx$$

can be simplified by substitution. Let $u = x - 50.0$. Then $du = dx$, and the limits of integration change accordingly:

When $x = 49.5$, $u = -0.5$ and when $x = 50.5$, $u = 0.5$. So,

$$\int_{49.5}^{50.5} 6(x-50.0)^2\,dx = 6\int_{-0.5}^{0.5} u^2\,du$$

The integral of u^2 is:

$$\begin{aligned}\int_{-0.5}^{0.5} u^2\,du &= \left[\frac{u^3}{3}\right]_{-0.5}^{0.5} \\ &= \frac{(0.5)^3}{3} - \frac{(-0.5)^3}{3} = \frac{0.125}{3} - \left(-\frac{0.125}{3}\right) \\ &= \frac{0.125}{3} + \frac{0.125}{3} \\ &= \frac{1}{12}\end{aligned}$$

and

$$6\int_{-0.5}^{0.5} u^2\,du = 6 \times \frac{1}{12} = \frac{6}{12} = 0.5$$

Therefore, from the Equation (5.2), we have

$$\int_{49.5}^{50.5} f(x)\,dx = 1.5 - 0.5 = 1$$

DRAFT

Since both conditions are satisfied, $f(x)$ is indeed a valid probability density function.

The graphical presentation of the density $f(x)$ is presented in Figure 5.6.

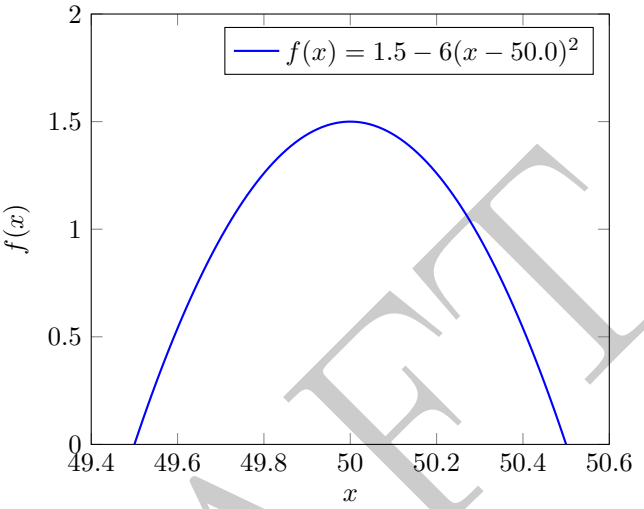


Figure 5.6: Density Plot of the PDF $f(x)$

(ii). We can find the probability that a metal cylinder has a diameter between 49.8 and 50.1 mm which is

$$\begin{aligned}
P(49.8 \le X \le 50.1) &= \int_{49.8}^{50.1} f(x)\,dx \\
&= \int_{49.8}^{50.1} \left(1.5 - 6(x - 50.0)^2\right)\,dx \\
&= \int_{49.8}^{50.1} 1.5\,dx - \int_{49.8}^{50.1} 6(x - 50.0)^2\,dx \\
&= 1.5\,[x]_{49.8}^{50.1} - 6\int_{-0.2}^{0.1} u^2\,du \quad [\text{let}, \quad u = x - 50.0] \\
&= 1.5(50.1 - 49.8) - 6\int_{-0.2}^{0.1} u^2\,du \\
&= 0.45 - \frac{(0.1)^3}{3} - \frac{(-0.2)^3}{3} \\
&= 0.45 - 0.018 = 0.432
\end{aligned}$$

Thus, the probability that a metal cylinder has a diameter between 49.8 and 50.1 mm is 0.432 or 43.2%.

DRAFT

5.4.2 Cumulative Distribution Function

The **cumulative distribution function (CDF)** of a continuous random variable X is a function that gives the probability that X will take a value less than or equal to x. The cumulative distribution function can be calculated from the expression:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy.$$

In practical applications, the lower integration limit of $-\infty$ can be replaced by the lower boundary of the state space, since the probability density function (PDF) is zero outside this region. The PDF can be obtained by differentiating the cumulative distribution function, which is given by:

$$f(x) = \frac{dF(x)}{dx}.$$

This relationship connects the probability density function to the cumulative probability.

Cumulative Distribution Function: The cumulative distribution function $F(x)$ of a random variable X is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy.$$

For a continuous random variable X, the three properties mentioned in Section 5.3.3 are satisfied. In addition, the following property must hold: if the CDF $F(\cdot)$ is continuous at any $a \leq x \leq b$, then

$$P(a \leq X \leq b) = F(b) - F(a).$$

Note that $F(x)$ is a non-decreasing function, meaning that if $a < b$, then $F(a) \leq F(b)$. This reflects that as you move to the right along the x-axis, the cumulative probability does not decrease. Another important property is as follows:

- $\lim_{x \to -\infty} F(x) = 0$: As x approaches negative infinity, the CDF approaches 0, indicating that the probability of the random variable being less than any finite value is 0.
- $\lim_{x \to \infty} F(x) = 1$: As x approaches positive infinity, the CDF approaches 1, indicating that the probability of the random variable being less than any sufficiently large value is 1.

Problem 5.4. *Let X be a continuous random variable with the probability density function:*

$$f(x) = \begin{cases} 2x & \textit{for } 0 \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

DRAFT

(i). Verify that the function $f(x)$ is a valid probability density function by showing that the total area under the curve is equal to 1.

(ii). Calculate the probability $P(0.5 < X < 0.8)$.

(iii). Derive the cumulative distribution function $F(x)$.

Solution

(i). To verify that the function

$$f(x) = \begin{cases} 2x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

is a valid probability density function, we calculate the integral over its range:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{1} 2x\,dx$$

Calculating the integral:

$$\int_{0}^{1} 2x\,dx = \left[x^2\right]_0^1 = 1^2 - 0^2 = 1$$

Since

$$\int_{-\infty}^{\infty} f(x)\,dx = 1,$$

we conclude that $f(x)$ is a valid probability density function.

(ii). We compute

$$P(0.5 < X < 0.8) = \int_{0.5}^{0.8} 2x\,dx = \left[x^2\right]_{0.5}^{0.8} = (0.8)^2 - (0.5)^2 = 0.39$$

Thus, $P(0.5 < X < 0.8) = 0.39$.

(iii). The cumulative distribution function $F(x)$ is given by:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

For $0 \leq x \leq 1$:

$$F(x) = \int_{0}^{x} 2t\,dt = \left[t^2\right]_0^x = x^2$$

So, the CDF $F(x)$ can be summarized as:

DRAFT

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^2 & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

□

Problem 5.5. *Consider the Problem 5.3, the probability density function $f(x)$ of a random variable X is*

$$f(x) = \begin{cases} 1.5 - 6(x - 50.0)^2 & \textit{for } 49.5 \le x \le 50.5, \\ 0 & \textit{otherwise.} \end{cases}$$

Find the cumulative distribution function of X and give the graphical presentation of this function.

Solution

The cumulative distribution function of X is defined as

$$\begin{aligned} F(x) &= \int_{49.5}^{x} (1.5 - 6(t - 50.0)^2)\, dt \\ &= \int_{49.5}^{x} 1.5\, dt - \int_{49.5}^{x} 6(t - 50.0)^2\, dt \\ &= 1.5(x - 49.5) - 6\int_{-0.5}^{x-50} u^2\, du \quad [\text{let}, \quad u = t - 50.0] \\ &= 1.5(x - 49.5) - 6\left[\frac{u^3}{3}\right]_{-0.5}^{x-50} \\ &= 1.5(x - 49.5) - 2\left((x - 50)^3 + 0.125\right) \\ &= 1.5(x - 49.5) - 2(x - 50)^3 - 0.25 \end{aligned}$$

Therefore, the CDF $F(x)$ is:

$$F(x) = \begin{cases} 0 & \text{for } x < 49.5 \\ 1.5(x - 49.5) - 2(x - 50)^3 - 0.25 & \text{for } 49.5 \le x \le 50.5 \\ 1 & \text{for } x > 50.5 \end{cases}$$

The graphical presentation of $F(x)$ is dipicted in Figure 5.7.

DRAFT

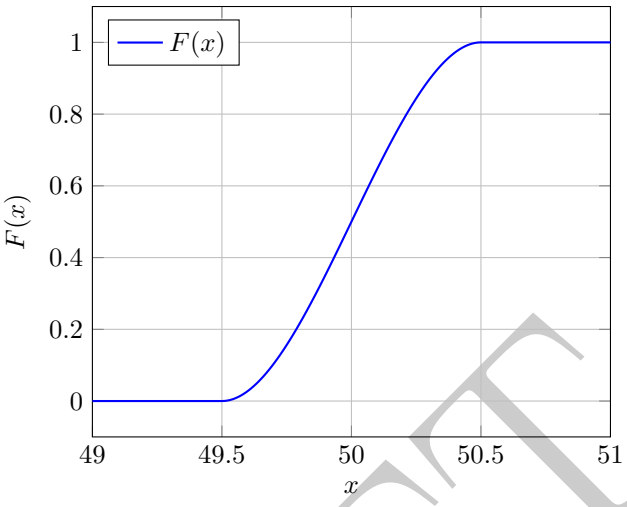


Figure 5.7: The cumulative distribution function of $f(x)$.

Problem 5.6. *Given the cumulative distribution function:*

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^3 & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

(i). Find the probability density function $f(x)$. Verify that the $f(x)$ is a valid PDF.

(ii). Calculate the probability $P(0.2 < X < 0.5)$.

Solution

(i). The PDF is the derivative of the CDF. That is

$$f(x) = \frac{d}{dx} F(x)$$

Taking the derivative for each region, we have

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } x > 1 \end{cases}$$

Verification that $f(x)$ is a valid PDF: It is noted that $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^1 3x^2\, dx = 3\left[\frac{x^3}{3}\right]_0^1 = 1.$$

DRAFT

Since $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\,dx = 1$, $f(x)$ is a valid probability density function.

(ii). To find the probability $P(0.2 < X < 0.5)$, we can use the PDF:

$$\begin{aligned} P(0.2 < X < 0.5) &= \int_{0.2}^{0.5} f(x)\,dx = \int_{0.2}^{0.5} 3x^2\,dx \\ &= \left[x^3\right]_{0.2}^{0.5} = (0.5)^3 - (0.2)^3 \\ &= 0.125 - 0.008 = 0.117 \end{aligned}$$

Alternatively,

$$P(0.2 < X < 0.5) = F(0.5) - F(0.2) = 0.5^3 - 0.2^3 = 0.117$$

Hence, the probability $P(0.2 < X < 0.5)$ is 0.117. □

Problem 5.7. *Let X be a continuous random variable with the PDF:*

$$f(x) = \begin{cases} \frac{1}{2}e^{-|x|} & \text{for } -\infty < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute $F(x)$.

Solution

For $-\infty < x < \infty$, the cumulative function is

$$F(x) = \int_{-\infty}^{x} \frac{1}{2}e^{-|t|}\,dt$$

Since $e^{-|t|}$ can be split into two parts depending on the range of t.
For $x < 0$:

$$F(x) = \int_{-\infty}^{x} \frac{1}{2}e^{t}\,dt = \frac{1}{2}\left[e^t\right]_{-\infty}^{x} = \frac{1}{2}\left(e^x - 0\right) = \frac{1}{2}e^x$$

For $x \geq 0$:

$$\begin{aligned} F(x) &= \int_{-\infty}^{0} \frac{1}{2}e^{t}\,dt + \int_{0}^{x} \frac{1}{2}e^{-t}\,dt \\ &= \frac{1}{2}\left[e^t\right]_{-\infty}^{0} + \frac{1}{2}\left[-e^{-t}\right]_{0}^{x} \\ &= \frac{1}{2}\left(1 - 0\right) + \frac{1}{2}\left(1 - e^{-x}\right) \\ &= 1 - \frac{1}{2}e^{-x} \end{aligned}$$

So, the CDF $F(x)$ is:

$$F(x) = \begin{cases} \frac{1}{2}e^{x} & \text{for } x < 0 \\ 1 - \frac{1}{2}e^{-x} & \text{for } x \geq 0. \end{cases}$$

□

DRAFT

5.4.3 Exercises

1. Consider a random variable measuring the following quantities. In each case, state with reasons whether you think it is more appropriate to define the random variable as discrete or continuous.

   (a) The number of books in a library

   (b) The duration of a phone call

   (c) The number of steps a person takes in a day

   (d) The amount of rainfall in a month

   (e) The number of languages a person speaks

   (f) The speed of a car on a highway

2. A random variable X takes values between 4 and 6 with a probability density function

$$f(x) = \frac{k}{x \ln(1.5)} \quad \text{for } 4 \leq x \leq 6.$$

   (a) What is the value of k?

   (b) Make a sketch of the probability density function.

   (c) What is $P(4.5 \leq X \leq 5.5)$?

   (d) Construct and sketch the cumulative distribution function.

3. A random variable Y takes values between 1 and 3 with a probability density function

$$g(y) = \frac{k}{(y+1)^2} \quad \text{for } 1 \leq y \leq 3.$$

   (a) Find the value of k and then make a sketch of the probability density function.

   (b) What is $P(1.5 \leq Y \leq 2.5)$?

   (c) Construct and sketch the cumulative distribution function.

4. A random variable Z takes values between 2 and 5 with a probability density function

$$h(z) = \frac{k}{(z+1)^3} \quad \text{for } 2 \leq z \leq 5.$$

   (a) Find the value of k and then make a sketch of the probability density function.

   (b) What is $P(2.5 \leq Z \leq 4)$?

DRAFT

(c) Construct and sketch the cumulative distribution function.

5. A random variable X takes values between 0 and 4 with a cumulative distribution function

$$F(x) = \frac{x^2}{16} \quad \text{for } 0 \leq x \leq 4.$$

(a) Sketch the cumulative distribution function.

(b) What is $P(X \leq 2)$?

(c) What is $P(1 \leq X \leq 3)$?

(d) Construct and sketch the probability density function.

6. The resistance X of an electrical component has a probability density function

$$f(x) = Ax(130 - x^2) \quad \text{for resistance values in the range } 10 \leq x \leq 11.$$

(a) Calculate the value of the constant A.

(b) Calculate the cumulative distribution function.

(c) What is the probability that the electrical component has a resistance between 10.25 and 10.5?

5.5 The Expectation of a Random Variable

While the probability mass function or the probability density function provides complete information about the probabilistic properties of a random variable, it is often useful to use some summary measures of these properties. One of the most fundamental summary measures is the expectation or mean of a random variable, denoted by $E(X)$, which represents the "average" value of the random variable. Two random variables with the same expected value can be considered to have the same average value, even though their probability mass functions or probability density functions may differ significantly.

The **expected value** (or mean) of a random variable is a measure of its central tendency.

Expected Value of a Random Variable: For a discrete random variable X,

$$E[X] = \sum_x x \cdot P(X = x)$$

DRAFT

For a continuous random variable X,

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

5.5.1 Example: Testing Electronic Components

To find the expectation (expected value) $E[X]$ for the random variable X representing the number of defective components in the three tested electronic components, we use the definition of the expectation for a discrete random variable:

$$E[X] = \sum_{x} x \cdot P(X = x)$$

Given the probability mass function for X in Table 5.2, the calculations for the $E[X]$ are shown in the following table.

x	$P(X = x)$	$x \cdot P(X = x)$
0	$\frac{1}{8}$	0
1	$\frac{3}{8}$	$1 \cdot \frac{3}{8} = \frac{3}{8}$
2	$\frac{3}{8}$	$2 \cdot \frac{3}{8} = \frac{6}{8}$
3	$\frac{1}{8}$	$3 \cdot \frac{1}{8} = \frac{3}{8}$
Total		$\frac{12}{8}$

Using the output in the above table

$$E[X] = \sum_{x} x \cdot P(X = x) = \frac{12}{8} = 1.5$$

Therefore, the expected number of defective components $E[X]$ is 1.5.

Problem 5.8.

An office has four copying machines, and the random variable X denotes how many of them are in use in a particular time. Suppose the probability mass function X is given below:

x	0	1	2	3	4
$\Pr(X = x)$	k	0.02	0.05	0.4	$(k + 0.2)$

(a) What is the value of k.

(b) Find the expectation of X.

DRAFT

Solution

(a) To find the value of k, we set up the equation based on the property that the sum of probabilities must equal 1:

$$\text{or, } k + 0.02 + 0.05 + 0.4 + (k + 0.2) = 1$$

Solving for k:

$$\text{or, } 2k + 0.67 = 1$$

$$\text{or, } 2k = 1 - 0.67$$

$$\text{or, } 2k = 0.33$$

$$\therefore k = \frac{0.33}{2} = 0.165$$

(b) To find the expectation $E(X)$ of the random variable X, we use the formula for the expected value:

$$E(X) = \sum_{x=0}^{4} x \cdot \Pr(X = x)$$

Given the probability mass function:

$$\begin{aligned}
&\Pr(X = 0) = k = 0.165,\\
&\Pr(X = 1) = 0.02,\\
&\Pr(X = 2) = 0.05,\\
&\Pr(X = 3) = 0.4,\\
&\Pr(X = 4) = k + 0.2 = 0.165 + 0.2 = 0.365.
\end{aligned}$$

Now we can calculate $E(X)$:

$$E(X) = 0{\cdot}\Pr(X = 0){+}1{\cdot}\Pr(X = 1){+}2{\cdot}\Pr(X = 2){+}3{\cdot}\Pr(X = 3){+}4{\cdot}\Pr(X = 4)$$

Substituting the values:

$$\begin{aligned}
E(X) &= 0 \cdot 0.165 + 1 \cdot 0.02 + 2 \cdot 0.05 + 3 \cdot 0.4 + 4 \cdot 0.365\\
&= 0 + 0.02 + 0.1 + 1.2 + 1.46\\
&= 0.02 + 0.1 + 1.2 + 1.46\\
&= 2.78.
\end{aligned}$$

Thus, the expectation of X is 2.78.

DRAFT

5.5.2 Example: Metal Cylinder Production

The probability density function of the diameter of a metal cylinder (X) is

$$f(x) = 1.5 - 6(x - 50.0)^2 \quad \text{for } 49.5 \leq x \leq 50.5$$

The expectation $E(X)$ is calculated as follows:

$$\begin{aligned}
E(X) &= \int_{49.5}^{50.5} x f(x)\, dx = \int_{49.5}^{50.5} x\left(1.5 - 6(x-50.0)^2\right)\, dx \\
&= \int_{49.5}^{50.5} 1.5x\, dx - \int_{49.5}^{50.5} 6x(x-50)^2\, dx \\
&= 1.5\left[\frac{x^2}{2}\right]_{49.5}^{50.5} - \int_{49.5}^{50.5} 6x(x-50)^2\, dx \quad [\text{let}, \quad u = x - 50] \\
&= 1.5\left(\frac{50.5^2}{2} - \frac{49.5^2}{2}\right) - 6\int_{-0.5}^{0.5} (u^3 + 50u^2)\, du \\
&= 1.5\left(\frac{2550.25}{2} - \frac{2450.25}{2}\right) - 6\int_{-0.5}^{0.5} u^3\, du - 6\left(50\left[\frac{u^3}{3}\right]_{-0.5}^{0.5}\right) \\
&= 75 - 0 - 6 \times 50 \times \left(\frac{0.5^3}{3} + \frac{0.5^3}{3}\right) \\
&= 75 - 1.5 \times 50 \\
&= 50
\end{aligned}$$

Hence the expectation of the diameter of a metal cylinder is 50mm.

5.5.3 Exercises

1. Suppose the Laptop repair costs C (in dollars) are \$50, \$200, and \$350 with respective probability values of 0.3, 0.2, and 0.5.
   (a) Write the probability mass function for the repair cost C.
   (b) Verify that this is a valid probability distribution.
   (c) Find the cumulative distribution function $F(c)$.
   (d) Find the probability that the repair cost is at most \$200.
   (e) Calculate the expected Laptop repair cost.
2. Let X be the number of defective items in a batch of 5 products with probability mass function
$$P(X = x) = \frac{x+1}{15} \qquad x = 0, 1, 2, 3, 4.$$
   (a) Verify that this is a valid probability mass function.

DRAFT

(b) Find the cumulative distribution function $F(x)$.

(c) Find the probability that there are at most 2 defective items.

(d) Find the expected number of defective items.

(e) Find the variance of the number of defective items.

3. The lifetime X (in years) of a certain type of light bulb is either 1, 3, or 5 with respective probability values of 0.2, 0.5, and 0.3.

(a) Write the probability mass function for X.

(b) Verify that this is a valid probability distribution.

(c) Find the cumulative distribution function $F(x)$.

(d) Find the probability that the lifetime is at least 3 years.

(e) Calculate the expected lifetime of the light bulb.

4. Let the temperature X in degrees Fahrenheit of a particular chemical reaction with density

$$f(x) = \frac{x - 190}{3600} \qquad 220 \leq x \leq 280.$$

Find the expectation of the temperature.

5. The weight W (in grams) of a manufactured part has probability density function

$$f(w) = \frac{w - 100}{1800} \qquad 120 \leq w \leq 180.$$

(a) Find the probability that the weight is more than 150 grams.

(b) Find the cumulative distribution function $F(w)$ for $120 \leq w \leq 180$.

(c) Find the expected weight of the part.

6. The time T (in hours) between successive arrivals of customers at a service center follows the following probability density function

$$f(t) = \lambda e^{-\lambda t} \qquad t \geq 0,$$

where $\lambda = 0.5$ customers per hour.

(a) Find the probability that the time between two successive arrivals is more than 3 hours.

(b) Find the cumulative distribution function $F(t)$ for $t \geq 0$.

(c) Find the expected time between successive customer arrivals.

7. The reaction time R (in seconds) of a subject in an experiment has probability density function

$$f(r) = \frac{3r^2}{26} \qquad 1 \leq r \leq 3.$$

DRAFT

(a) Find the probability that the reaction time is less than 2 seconds.

(b) Find the cumulative distribution function $F(r)$ for $1 \leq r \leq 3$.

(c) Find the expected reaction time.

5.6 The Variance of a Random Variable

Another key summary measure of the distribution of a random variable is the variance, which quantifies the spread or variability in the values that the random variable can take. While the mean or expectation captures the central or average value of the random variable, the variance measures the dispersion or deviation of the random variable around its mean value. Specifically, the variance of a random variable is defined as

$$\text{Var}(X) = E((X - E(X))^2).$$

This means that the variance is the expected value of the squared deviations of the random variable values from the expected value $E(X)$. The variance is always positive, and larger values indicate a greater spread in the distribution of the random variable around the mean. An alternative and often simpler expression for calculating the variance is

$$\begin{aligned}\text{Var}(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2.\end{aligned}$$

Variance: The variance of a random variable X is defined as

$$\text{Var}(X) = E((X - E(X))^2)$$

or equivalently

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

The variance is a positive measure that indicates the spread of the distribution of the random variable around its mean value. Larger values of the variance suggest that the distribution is more spread out.

It is typical to use the symbol μ to represent the mean or expectation of a random variable, and the symbol σ^2 to represent the variance. The standard deviation, denoted by σ, is the square root of the variance and is often used instead of the variance to describe the spread of the distribution.

DRAFT

Standard Deviation: The standard deviation of a random variable X is defined as the positive square root of the variance. The variance of a random variable is commonly denoted by σ^2, so σ represents the standard deviation.

The concept of variance can be illustrated graphically. Figure 5.8 shows two probability density functions with different mean values but identical variances. The variances are the same because the shape or spread of the density functions around their mean values is the same. In contrast, Figure 5.9 shows two probability density functions with the same mean values but different variances. The density function that is flatter and more spread out has the larger variance.

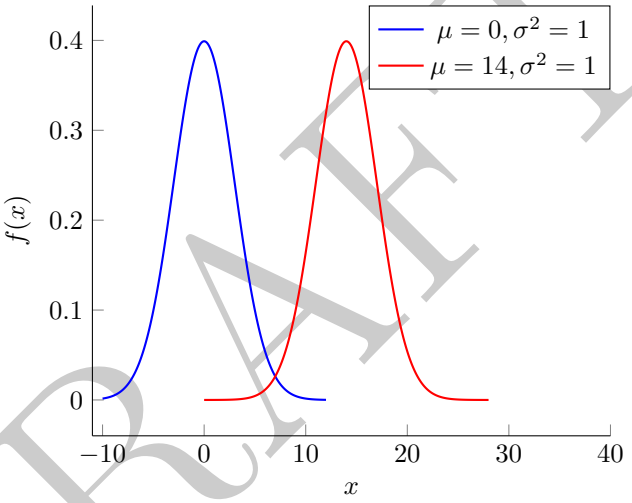


Figure 5.8: Two normal distributions with different means but identical variances.

DRAFT

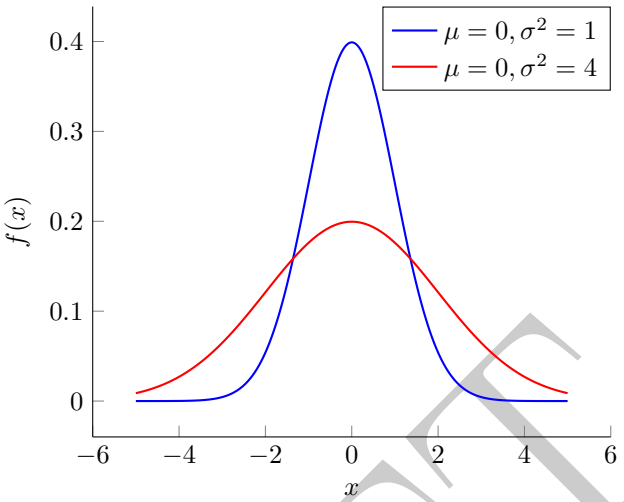


Figure 5.9: Two normal distributions with identical means but different variances.

It is important to note that the standard deviation has the same units as the random variable X, while the variance has units that are squared. For instance, if the random variable X is measured in seconds, then the standard deviation will also be in seconds, but the variance will be measured in seconds squared (seconds^2).

Problem 5.9. *Suppose a factory inspects small batches of 3 items for defects. Let the random variable Y denote the number of defective items in a batch. The probability mass function of Y is given below:*

y	$P(Y = y)$
0	$\frac{1}{5}$
1	$\frac{2}{5}$
2	$\frac{1}{4}$
3	$\frac{1}{5}$

Tasks:

(i). Compute the expected value $E[Y]$

(ii). Compute $E[Y^2]$

(iii). Find the variance of Y

(iv). Find the standard deviation of Y

DRAFT

Solution: We begin by calculating the necessary values in the following table:

Table 5.5: Calculating $E[Y]$ and $E[Y^2]$

y	$P(Y=y)$	$y \cdot P(Y=y)$	$y^2 \cdot P(Y=y)$
0	$\frac{1}{5}$	$0 \cdot \frac{1}{5} = 0$	$0^2 \cdot \frac{1}{5} = 0$
1	$\frac{2}{5}$	$1 \cdot \frac{2}{5} = \frac{2}{5}$	$1^2 \cdot \frac{2}{5} = \frac{2}{5}$
2	$\frac{1}{4}$	$2 \cdot \frac{1}{4} = \frac{2}{4}$	$2^2 \cdot \frac{1}{4} = \frac{4}{4}$
3	$\frac{1}{5}$	$3 \cdot \frac{1}{5} = \frac{3}{5}$	$3^2 \cdot \frac{1}{5} = \frac{9}{5}$
Total		$\frac{2}{5} + \frac{2}{4} + \frac{3}{5} = \frac{30}{20} = 1.5$	$\frac{2}{5} + 1 + \frac{9}{5} = \frac{16}{5} = 3.2$

(i). Expected Value

$$E[Y] = 1.5$$

(ii). Second Moment

$$E[Y^2] = 3.2$$

(iii). Variance

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = 3.2 - (1.5)^2 = 3.2 - 2.25 = 0.95$$

(iv). Standard Deviation

$$\sigma_Y = \sqrt{\text{Var}(Y)} = \sqrt{0.95} \approx 0.9747$$

□

5.6.1 Example: Metal Cylinder Production

The probability density function of the diameter of a metal cylinder (X) is

$$f(x) = 1.5 - 6(x - 50.0)^2 \quad \text{for } 49.5 \leq x \leq 50.5$$

DRAFT

and the $E(X) = 50$. To find the variance $V(X)$, we need $E(X^2)$:

$$\begin{aligned}
E(X^2) &= \int_{49.5}^{50.5} x^2 f(x)\,dx = \int_{49.5}^{50.5} x^2 \left(1.5 - 6(x-50.0)^2\right)\,dx \\
&= \int_{49.5}^{50.5} 1.5x^2\,dx - \int_{49.5}^{50.5} 6x^2(x-50)^2\,dx \\
&= 1.5\left[\frac{x^3}{3}\right]_{49.5}^{50.5} - 6\int_{-0.5}^{0.5} (u+50)^2 u^2\,du \quad [\text{let}, \quad u = x - 50] \\
&= 1.5\left(\frac{50.5^3}{3} - \frac{49.5^3}{3}\right) - 6\int_{-0.5}^{0.5} (u^4 + 100u^2 + 2500)\,du \\
&= 2500.05 \quad [\text{after simplification}]
\end{aligned}$$

Therefore, the variance is

$$V(X) = E(X^2) - (E(X))^2 = 2500.05 - 2500 = 0.05$$

Thus, the variance $V(X) = 0.05$ and the standard deviation $\sigma = \sqrt{0.05} = 0.2236$.

Problem 5.10. *Consider a random variable X representing the number of heads in three tosses of a fair coin. The possible values of X are 0, 1, 2, and 3. The PMF of X is given by:*

$$P(X = x) = \binom{3}{x}\left(\frac{1}{2}\right)^3$$

Find the expected value and variance of X.

5.6.2 Chebyshev's Inequality

When we look at a collection of numbers, like test scores, people's heights, or daily temperatures, we often want to know how much they vary. We typically calculate two key values: the **average** (or mean, μ) and the **standard deviation** (σ), which tells us how much the numbers typically spread out from the average.

But what if we don't know the exact pattern or "shape" of how these numbers are distributed? For instance, are most scores tightly grouped around the average, or are they spread widely with many very high and very low scores?

This is where **Chebyshev's Inequality** becomes incredibly useful! It acts as a universal rule that tells us a minimum percentage of our data will be found close to the average, regardless of the data's specific distribution. This makes it a powerful tool in data science for tasks like quality control (ensuring most products meet standards) or salary analysis (understanding the typical range

DRAFT

of earnings).

Consider a set of numbers (represented by a "random variable" X) with:

- Its average (mean): μ
- Its typical spread (standard deviation): σ

Chebyshev's Inequality provides a boundary. For *any* positive number $k > 1$, we can state the inequality in two equivalent forms:

- **Form 1 (Probability of being far):**

$$P\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

  This means: "The probability that a number X deviates from its average (μ) by k or more standard deviations is at most $1/k^2$."

- **Form 2 (Probability of being close):**

$$P\left(\mu - k\sigma \leq X \leq \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

  This means: "The probability that a number X falls within k standard deviations of the average (from $\mu - k\sigma$ to $\mu + k\sigma$) is at least $1 - 1/k^2$."

Note that these two forms are complements of each other, since the probability of being close plus the probability of being far must equal 1.

Chebyshev's Inequality: Let X be a random variable with mean μ and variance σ^2. For any $k > 1$, the probability that X falls within k standard deviations of the mean is at least $1 - 1/k^2$:

$$P\left(\mu - k\sigma \leq X \leq \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}.$$

Equivalently, the probability that X deviates from the mean by k or more standard deviations is at most $1/k^2$:

$$P\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

This inequality demonstrates that the further you move away from the average (as k increases), the less likely it is to find data points. Let's look at some examples:

- If you choose $k = 2$: At least $1 - 1/2^2 = 1 - 1/4 = 3/4$ (or 75%) of your data will be found within 2 standard deviations of the average.

DRAFT

- If you choose $k = 3$: At least $1-1/3^2 = 1-1/9 = 8/9$ (or approximately 89%) of your data will be found within 3 standard deviations of the average.

Problem 5.11. *A major online retailer advertises that its average delivery time for standard shipping is 5 days, with a standard deviation of 1.5 days. Assuming nothing is known about the shape of the delivery time distribution, what can we say about the percentage of deliveries that will take:*

(i). between 2 days and 8 days?

(ii). more than 9 days, or less than 1 day?

Solution: We can use Chebyshev's Inequality to answer these questions, as we are given the mean and standard deviation but nothing about the distribution's specific shape. Let X be the random variable representing the delivery time in days.
Given:

- Mean (μ) = 5 days
- Standard deviation (σ) = 1.5 days

(i). Percentage of deliveries between 2 days and 8 days: We are looking for the probability $P(2 \leq X \leq 8)$. First, we need to express the range $[2, 8]$ in terms of $\mu \pm k\sigma$. The center of this range is 5, which is our mean. The distance from the mean to either end of the interval is $8 - 5 = 3$ days (or $5 - 2 = 3$ days). So, we have a deviation of 3 days from the mean. We set this deviation equal to $k\sigma$:

$$\begin{aligned} k\sigma &= 3 \\ k(1.5) &= 3 \\ k &= \frac{3}{1.5} = 2 \end{aligned}$$

Now, apply Chebyshev's Inequality for $k = 2$:

$$P\left(\mu - k\sigma \leq X \leq \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Substitute the values:

$$P\left(5 - 2(1.5) \leq X \leq 5 + 2(1.5)\right) \geq 1 - \frac{1}{2^2}$$

$$P\left(5 - 3 \leq X \leq 5 + 3\right) \geq 1 - \frac{1}{4}$$

$$P\left(2 \leq X \leq 8\right) \geq \frac{3}{4}$$

DRAFT

$$P\left(2 \leq X \leq 8\right) \geq 0.75$$

Interpretation At least 75% of all deliveries will take between 2 and 8 days.

(ii). Percentage of deliveries taking more than 9 days, or less than 1 day: We are looking for the probability $P(X > 9 \text{ or } X < 1)$. This is equivalent to $P(|X - \mu| \geq \text{deviation})$. The deviations from the mean (5 days) are: $9 - 5 = 4$ days $5 - 1 = 4$ days So, the deviation is 4 days. We set this deviation equal to $k\sigma$:

$$\begin{aligned} k\sigma &= 4 \\ k(1.5) &= 4 \\ k &= \frac{4}{1.5} = \frac{8}{3} \approx 2.67 \end{aligned}$$

Now, apply Chebyshev's Inequality for $k = \frac{8}{3}$:

$$P\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

Substitute the values:

$$P\left(|X - 5| \geq \frac{8}{3}(1.5)\right) \leq \frac{1}{\left(\frac{8}{3}\right)^2}$$

$$P\left(|X - 5| \geq 4\right) \leq \frac{1}{\frac{64}{9}}$$

$$P\left(|X - 5| \geq 4\right) \leq \frac{9}{64}$$

$$P\left(X < 1 \text{ or } X > 9\right) \leq 0.140625$$

Interpretation At most approximately 14.06% of deliveries will take less than 1 day or more than 9 days. This also implies that at least $100\% - 14.06\% = 85.94\%$ of deliveries will fall between 1 and 9 days.

Problem 5.12. *A statistics exam has scores with a mean $\mu = 70$ and a standard deviation $\sigma = 8$. Use Chebyshev's Inequality to find the range around the mean within which* ***at least 89%*** *of the exam scores fall.*

Solution: Chebyshev's Inequality states that for any random variable X with mean μ and standard deviation σ:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

We want to find k such that:

DRAFT

$$1 - \frac{1}{k^2} \geq 0.89$$

Rearranging, we get:

$$\frac{1}{k^2} \leq 0.11$$

$$k^2 \geq \frac{1}{0.11} \approx 9.09$$

$$k \geq \sqrt{9.09} \approx 3.02$$

Next, calculate the interval around the mean:

$$\mu \pm k\sigma = 70 \pm 3.02 \times 8 = 70 \pm 24.16$$

Thus, the range is:

$$[70 - 24.16,\ 70 + 24.16] = [45.84,\ 94.16]$$

At least 89% of exam scores lie between 45.84 and 94.16. □

Problem 5.13. *A company has an average salary of \$60,000 with a standard deviation of \$5,000. Use Chebyshev's Inequality to determine the salary range within which* ***at least 80%*** *of the employees' salaries fall.*

Solution: By Chebyshev's Inequality, for a random variable X with mean μ and standard deviation σ, the probability that X lies within k standard deviations of the mean satisfies:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

We require this probability to be at least 0.80:

$$1 - \frac{1}{k^2} \geq 0.80 \implies \frac{1}{k^2} \leq 0.20.$$

Solving for k, we obtain:

$$k^2 \geq 5 \implies k \geq \sqrt{5} \approx 2.24.$$

Hence, the range containing at least 80% of salaries is:

$$60,000 \pm 2.24 \times 5,000 = 60,000 \pm 11,200,$$

or equivalently,

$$[48,800,\quad 71,200].$$

Therefore, at least 80% of employees earn between \$48,800 and \$71,200.

DRAFT

5.6.3 Quantiles of Random Variables

Quantiles are useful summary measures that provide insight into the spread or variability of a random variable's distribution. The p-th quantile of a random variable X, which has a cumulative distribution function $F(x)$, is the value x that satisfies

$$F(x) = p$$

meaning that there is a probability p that the random variable is less than the p-th quantile. The probability p is often expressed as a percentage, and the corresponding quantiles are known as percentiles. For instance, the 70th percentile is the value x for which $F(x) = 0.70$. It is important to note that the 50th percentile of a distribution is also known as the median.

Quantiles: The p-th quantile of a random variable X with a cumulative distribution function $F(x)$ is the value x such that

$$F(x) = p$$

This is also known as the $p \times 100$-th percentile of the random variable. The probability p signifies the chance that the random variable takes on a value less than the p-th quantile.

To understand the spread of a distribution, one can compute its quartiles. The upper quartile is the 75th percentile, and the lower quartile is the 25th percentile. Together with the median, these quartiles divide the range of the random variable into four equal parts, each with a probability of 0.25.

The interquartile range, which is the distance between the upper and lower quartiles, serves as an indicator of distribution spread similar to variance. A larger interquartile range suggests that the distribution of the random variable is more spread out.

Quartiles and Interquartile Range: The *lower quartile* (Q_1) of a distribution is the value below which 25% of the data fall, also known as the 25th percentile. Similarly, the *upper quartile* (Q_3) is the value below which 75% of the data fall, or the 75th percentile. The *interquartile range* (IQR), defined as the difference between the upper and lower quartiles,

$$\text{IQR} = Q_3 - Q_1,$$

measures the spread of the middle 50% of the data. It is a robust measure of variability that is less sensitive to outliers than variance or standard deviation.

Problem 5.14. *The diameters X of metal cylinders produced by a factory follow a distribution with cumulative distribution function:*

$$F(x) = 1.5x - 2(x - 50.0)^3 - 74.5,$$

valid for $49.5 \leq x \leq 50.5$.

(i). Find the ***upper quartile*** Q_3.

(ii). Find the ***lower quartile*** Q_1.

(iii). Calculate the ***interquartile range (IQR)***.

Solution: Given the cumulative distribution function for the diameters X of metal cylinders:

$$F(x) = 1.5x - 2(x - 50.0)^3 - 74.5,$$

defined for $49.5 \leq x \leq 50.5$, we proceed as follows:

(i) **Find the upper quartile Q_3:**

By definition, the upper quartile Q_3 satisfies:

$$F(Q_3) = 0.75.$$

Substituting $F(x)$:

$$1.5Q_3 - 2(Q_3 - 50.0)^3 - 74.5 = 0.75.$$

Rearranging,

$$1.5Q_3 - 2(Q_3 - 50.0)^3 = 75.25.$$

This is a nonlinear equation in Q_3 and is solved numerically. The solution is approximately:

$$Q_3 \approx 50.17 \text{ mm}.$$

(ii) **Find the lower quartile Q_1:**

Similarly, the lower quartile Q_1 satisfies:

$$F(Q_1) = 0.25,$$

which gives:

$$1.5Q_1 - 2(Q_1 - 50.0)^3 - 74.5 = 0.25,$$

or equivalently,

DRAFT

$$1.5Q_1 - 2(Q_1 - 50.0)^3 = 74.75.$$

Solving numerically yields:

$$Q_1 \approx 49.83 \text{ mm}.$$

(iii) **Calculate the interquartile range (IQR):**

The IQR is defined as the difference between the upper and lower quartiles:

$$\text{IQR} = Q_3 - Q_1 = 50.17 - 49.83 = 0.34 \text{ mm}.$$

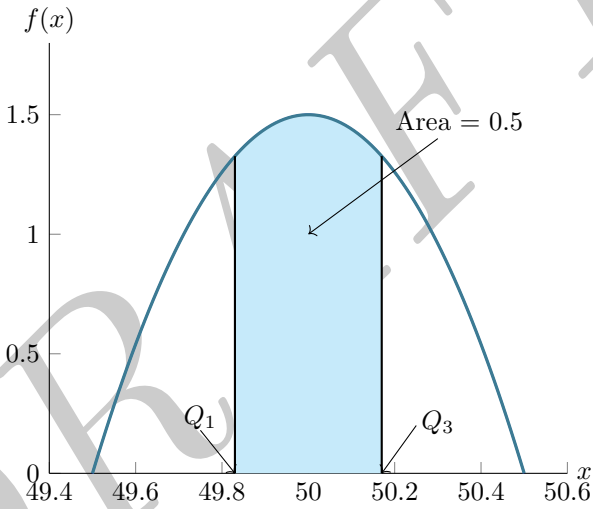


Figure 5.10: Interquartile range for metal cylinder diameters.

This range represents the middle 50% of cylinder diameters, indicating the variability of the manufacturing process. See Figure 5.10.

□

5.6.4 Linear Functions of a Random Variable

We will now explore some properties that will simplify calculating the means and variances of random variables discussed in later chapters. These properties allow us to express expectations using other parameters that are either known or easily computed. The results presented are applicable to both discrete and continuous random variables, though proofs are provided only for the continuous

DRAFT

case. We start with a theorem and two corollaries that should be intuitively understandable to the reader.

Theorem 5.1. *If a and b are constants, then*

$$E(aX + b) = aE(X) + b$$

and the variance is

$$Var(aX + b) = a^2\, Var(X).$$

Proof. By the definition of expected value,

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f(x)\, dx.$$

This can be rewritten as

$$E(aX + b) = a \int_{-\infty}^{\infty} x f(x)\, dx + b \int_{-\infty}^{\infty} f(x)\, dx.$$

The first integral on the right is $E(X)$ and the second integral equals 1. Therefore, we have

$$E(aX + b) = aE(X) + b.$$

Compute Var$(aX + b)$:

$$\begin{aligned}
\mathrm{Var}(aX + b) &= E[(aX + b - E(aX + b))^2] \\
&= E[(aX + b - (aE(X) + b))^2] \\
&= E[(aX + b - aE(X) - b)^2] \\
&= E[(aX - aE(X))^2] \\
&= E[a^2(X - E(X))^2] \\
&= a^2 E[(X - E(X))^2] \\
&a^2 \mathrm{Var}(X).
\end{aligned}$$

Thus, we have shown that:

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

□

Problem 5.15. *Applying Theorem 5.1 to the continuous random variable*

$$Y = 1.1X + 0.3,$$

rework Example 5.5.2.

DRAFT

For Example 5.5.2 and 5.6.1, it is obtained $E(X) = 50$ and $V(X) = 0.05$. We may use Theorem 5.1 to write

$$\begin{aligned} E[Y] &= 1.1E[X] + 0.3 \\ &= 1.1 \times 50 - 4.5 \\ &= 50.5 \end{aligned}$$

and

$$\text{Var}(Y) = (1.1)^2\text{Var}(X) = (1.1)^2 \times 0.05 = 0.0605$$

Problem 5.16. *Suppose that a temperature has a mean of* $110°F$ *and a standard deviation of* $2.2°F$*. The conversion formula from Fahrenheit to Centigrade is given by:*

$$F = \frac{9C}{5} + 32$$

where F *is the temperature in Fahrenheit and* C *is the temperature in Centigrade. What are the mean and the standard deviation in degrees Centigrade?*

Solution

To find the mean temperature in Centigrade, we use:

$$C_{\text{mean}} = \frac{5}{9}(F_{\text{mean}} - 32)$$

Substitute $F_{\text{mean}} = 110$:

$$C_{\text{mean}} = \frac{5}{9}(110 - 32) = \frac{5}{9} \times 78 = 43.\overline{3}^{\circ}\text{C}$$

To find the standard deviation in Centigrade, we use:

$$\sigma_C = \frac{5}{9}\sigma_F$$

Substitute $\sigma_F = 2.2$:

$$\sigma_C = \frac{5}{9} \times 2.2 = \frac{11}{9} \approx 1.22°\text{C}$$

Thus, the mean temperature is approximately $43.\overline{3}^{\circ}$C and the standard deviation is approximately 1.22°C.

Theorem 5.2. *The expected value of the sum or difference of two or more functions of a random variable* X *is the sum or difference of the expected values of the functions. That is,*

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)].$$

DRAFT

Problem 5.17. *Let X be a random variable with probability distribution as follows:*

x	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

Find the expected value of $Y = (X-1)^2$ and variance of X.

Solution

Applying Theorem 5.2 to the function $Y = (X-1)^2$, we can write

$$E[(X-1)^2] = E(X^2 - 2X + 1) = E(X^2) - 2E(X) + E(1).$$

From Theorem 5.1, $E(1) = 1$, and by direct computation,

$$E(X) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (2)(0) + (3)\left(\frac{1}{6}\right) = 1,$$

and

$$E(X^2) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (4)(0) + (9)\left(\frac{1}{6}\right) = 2.$$

Hence,

$$E[(X-1)^2] = 2 - (2)(1) + 1 = 1.$$

Now, to calculate the variance of Y, we need $E[Y^2]$. Since $Y = (X-1)^2$,

$$Y^2 = (X-1)^4.$$

We need to compute $E[(X-1)^4]$:

$$\begin{aligned} E[(X-1)^4] &= \sum_x (x-1)^4 f(x) \\ &= 0-1)^4 \cdot \frac{1}{3} + (1-1)^4 \cdot \frac{1}{2} + (2-1)^4 \cdot 0 + (3-1)^4 \cdot \frac{1}{6} \\ &= \frac{1}{3} + 0 + 0 + \frac{8}{3} \\ &= 3. \end{aligned}$$

Therefore,

$$\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2 = 3 - 1^2 = 2.$$

Problem 5.18. *The weekly demand for a particular drink, measured in thousands of liters, at a chain of convenience stores is a continuous random variable $g(X) = X^2 + X - 2$, where X has the following density function:*

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2, \\ 0, & \textit{elsewhere.} \end{cases}$$

DRAFT

Solution

To find the expected value of the weekly demand for the drink, we use Theorem 5.1:

$$E(X^2 + X - 2) = E(X^2) + E(X) - E(2).$$

From Theorem 5.1, $E(2) = 2$. By direct integration, we find:

$$E(X) = \int_1^2 2x(x-1)\,dx = \frac{5}{3},$$

and

$$E(X^2) = \int_1^2 2x^2(x-1)\,dx = \frac{17}{6}.$$

Thus,

$$E(X^2 + X - 2) = \frac{17}{6} + \frac{5}{3} - 2 = \frac{5}{2}.$$

Therefore, the average weekly demand for the drink at this chain of convenience stores is 2500 liters.

Example: Test Score Standardization

Suppose that the raw scores X from a particular testing procedure are distributed between -5 and 20 with an expected value of 10 and a variance of 7. In order to standardize the scores so that they lie between 0 and 100, the linear transformation

$$Y = 4X + 20$$

is applied to the scores. This means, for example, that a raw score of $x = 12$ corresponds to a standardized score of $y = (4 \times 12) + 20 = 68$.

The expected value of the standardized scores is then known to be

$$E(Y) = 4E(X) + 20 = (4 \times 10) + 20 = 60$$

with a variance of

$$\text{Var}(Y) = 4^2\text{Var}(X) = 4^2 \times 7 = 112$$

The standard deviation of the standardized scores is $\sigma_Y = \sqrt{112} = 10.58$, which is $4 \times \sigma_X = 4 \times \sqrt{7}$.

5.6.5 Exercises

1. Consider the Laptop repair costs discussed in question 1 of Exercises 5.5.3, calculate the variance and standard deviation of the number of copying machines in use at a particular moment.

DRAFT

2. In machine breakdown problem, suppose that electrical failures generally cost \$400 to repair, mechanical failures have repair cost of \$550, and operator misuse failures have an repair cost of only \$100. These repair costs generate a random variable cost, as illustrated in the following Table.

x_i	100	400	550
$\Pr(X = x_i)$	0.25	$2(k+1)$	0.4

   (a) What is the value of k?

   (b) Find the average of the repair costs.

   (c) Find the variance of the repair costs.

3. Let Z be the score obtained when rolling a biased die with probability mass function

$$P(Z = z) = \frac{z}{21} \qquad z = 1, 2, 3, 4, 5, 6.$$

   (a) Show that this is a valid probability distribution.

   (b) Find the probability of obtaining a score greater than 4.

   (c) Find the cumulative distribution function $F(z)$ for all values of z.

   (d) Find the expected score.

   (e) Find the standard deviation of the score.

4. A random variable X takes values between 4 and 6 with a probability density function

$$f(x) = \frac{1}{x \ln(1.5)} \quad \text{for } 4 \leq x \leq 6.$$

   (a) What is the variance of this random variable?

   (b) What is the standard deviation of this random variable?

   (c) Find the upper and lower quartiles of this random variable.

   (d) What is the interquartile range?

5. A random variable X represents the time (in hours) until failure of a certain machine part, which is uniformly distributed between 100 and 200 hours.

$$f(x) = \frac{1}{100} \quad \text{for } 100 \leq x \leq 200.$$

   (a) What is the variance of this random variable?

   (b) What is the standard deviation of this random variable?

   (c) Find the upper and lower quartiles of this random variable.

DRAFT

(d) What is the interquartile range?

6. A random variable X takes values between 0 and 4 with a cumulative distribution function of

$$F(x) = \frac{x^2}{16} \quad \text{for } 0 \leq x \leq 4.$$

(a) What is the variance of this random variable?

(b) What is the standard deviation of this random variable?

(c) Find the upper and lower quartiles of this random variable.

(d) What is the interquartile range?

7. Let the temperature X in degrees Fahrenheit of a particular chemical reaction with density

$$f(x) = \frac{x - 190}{3600} \qquad 220 \leq x \leq 280.$$

(a) Verify that $f(x)$ is a valid probability density function.

(b) Find the cumulative distribution function $F(x)$ for $220 \leq x \leq 280$.

(c) Find the probability that the temperature is between 240 and 260 degrees Fahrenheit.

(d) Find the expected temperature of the chemical reaction.

(e) Find the variance of the temperature.

(f) What is the interquartile range of this random variable?

8. A researcher is studying the cholesterol levels of a population of adults. The cholesterol levels are known to have a mean of 200 mg/dL and a standard deviation of 25 mg/dL.

(a) Using Chebyshev's Inequality, determine the minimum percentage of adults whose cholesterol levels are within 50 mg/dL of the mean.

(b) To guarantee that at least 85% of the population has cholesterol levels within a certain number of standard deviations from the mean, how many standard deviations from the mean are required?

(c) If a randomly selected adult has a cholesterol level of 250 mg/dL, what is the maximum probability that this level deviates from the mean by at least 50 mg/dL according to Chebyshev's Inequality?

9. Suppose that the battery failure time, measured in hours, has a probability density function given

$$f(x) = \frac{2}{(x+1)^3}$$

for $x \geq 0$.

DRAFT

(a) Find the expected battery failure time.

(b) What is the probability that the battery fails within the first 4 hours.

(c) Find the cumulative distribution function of the battery failure times.

(d) Find the median of the battery failure times.

10. Suppose that a random variable X has mean $E(X) = 12$ and variance $\text{Var}(X) = 9$. Solve the following problems.

    (a) $E(3X + 5)$;

    (b) $E(7 - 2X)$;

    (c) $\text{Var}(3X + 5)$;

    (d) $\text{Var}(7 - 2X)$.

11. The daily demand for a product has mean 80 units and variance 36. If the selling price is \$5 per unit and the daily fixed cost is \$120, then the daily profit is

$$P = 5X - 120.$$

    Find the mean and variance of the daily profit.

12. The body weight of a group of students has mean 65 kg and standard deviation 8 kg. Let

$$Y = 2.2046X$$

    denote the corresponding weight in pounds. Then solve the following problems.

    (a) the mean weight in pounds;

    (b) the standard deviation in pounds.

5.7 Essential Generating Functions

In probability theory and statistics, generating functions are a powerful tool used to analyze and manipulate probability distributions. There are three main types of generating functions:

- Moment Generating Function (MGF): Used for both discrete and continuous random variables.
- Probability Generating Function (PGF): Used for discrete random variables.
- Characteristic Function (CF): It complements other generating functions, such as the MGF and PGF, and is closely related to the Fourier transform.

DRAFT

5.7.1 Moment Generating Function

The Moment Generating Function (MGF) is a powerful tool in probability theory and statistics used to summarize the properties of a probability distribution. It can help us find moments (for example, mean, variance), combine variables, and understand the distribution better. The Moment Generating Function of a random variable X is defined as:

$$M(t) = E[e^{tX}].$$

5.7.2 Key Properties of MGF

- **Higher Order Moments**:
  - The first moment (**mean**) is found by taking the first derivative of the MGF and evaluating it at $t = 0$:
    $$\mu = E[X] = M'(0)$$
  - The second moment is found by taking the second derivative of the MGF and evaluating it at $t = 0$:
    $$E[X^2] = M''(0)$$
  - The rth moment is found by taking the rth derivative of the MGF and evaluating it at $t = 0$:
    $$E[X^r] = \frac{d^r M(t)}{dt^r}\bigg|_{t=0}$$
    Evaluate the Derivative at $t = 0$ After differentiating the MGF the required number of times, substitute $t = 0$ to obtain the moment.
  - The **variance** can be found by taking the second derivative of the MGF, evaluating at $t = 0$, and then using it with the mean:
    $$\text{Var}(X) = M''(0) - [M'(0)]^2$$
- **Combining Variables**:
  - If X and Y are independent random variables, the MGF of their sum $X + Y$ is the product of their individual MGFs:
    $$M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

Problem 5.19 (**Discrete Random Variable**). *Consider a discrete random variable X that takes values 1 1 with probability p and the value 0 with probability $1 - p$. Find the moment generating function of X, and hence find mean and variance.*

DRAFT

Solution

The MGF of a random variable X is given by:

$$M(t) = E[e^{tX}] = \sum_x e^{tx} P(X = x)$$

We have,

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p$$

Substituting into the MGF formula:

$$\begin{aligned} M(t) &= e^{t\cdot 1} P(X = 1) + e^{t\cdot 0} P(X = 0) \\ &= e^t \cdot p + e^0 \cdot (1 - p) \\ &= pe^t + (1 - p) \end{aligned}$$

Mean of X: To find the mean $E[X]$, we differentiate the MGF with respect to t and evaluate at $t = 0$:

$$M'(t) = \frac{d}{dt}\left(pe^t + (1 - p)\right) = pe^t$$

Evaluating at $t = 0$:

$$M'(0) = pe^0 = p$$

Thus, the mean of X is:

$$E[X] = p$$

Variance of X:
To find the variance, we first calculate the second moment $E[X^2]$, which is the second derivative of the MGF evaluated at $t = 0$:

$$M''(t) = \frac{d}{dt}\left(pe^t\right) = pe^t$$

Evaluating at $t = 0$

$$M''(0) = pe^0 = p$$

Thus, the second moment is

$$E[X^2] = p$$

The variance is given by

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= p - p^2 = p(1 - p) \end{aligned}$$

Thus, the variance of X is

$$\text{Var}(X) = p(1 - p)$$

DRAFT

Problem 5.20 (**Continuous Random Variable**). *The PDF of X is given by:*

$$f(x) = \begin{cases} 1 & \text{if} \quad 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

This is a density of the uniform distribution on $[0, 1]$. Find the moment generating function of X, and hence find the mean and variance.

Solution

For this density the moment generating function is

$$M(t) = E\left[e^{tX}\right] = \int_0^1 e^{tx} \cdot 1 \, dx = \left[\frac{e^{tx}}{t}\right]_0^1 = \frac{e^t - 1}{t}, \quad \text{for } t \neq 0.$$

For $t = 0$, $M(0) = 1$ (since $M(t)$ is always 1 at $t = 0$ for any distribution). The first derivative of $M(t)$ with respect to t is:

$$M'(t) = \frac{d}{dt}\left(\frac{e^t - 1}{t}\right).$$

Applying the quotient rule:

$$M'(t) = \frac{t \cdot e^t - (e^t - 1)}{t^2} = \frac{te^t - e^t + 1}{t^2} = \frac{e^t(t-1) + 1}{t^2}.$$

To evaluate the mean, we need to find the limit of $M'(t)$ as $t \to 0$. We have the indeterminate form $\frac{0}{0}$, so we apply L'Hôpital's rule. To do this, we need to differentiate the numerator and denominator separately. So, we find,

$$\frac{d}{dt}\left((t-1)e^t + 1\right) = \frac{d}{dt}\left((t-1)e^t\right) + \frac{d}{dt}(1) = e^t(t-1) + e^t = e^t(t).$$

and

$$\frac{d}{dt}(t^2) = 2t.$$

So, applying L'Hôpital's rule, we find

$$M'(0) = \lim_{t\to 0} \frac{(t-1)e^t + 1}{t^2} = \lim_{t\to 0} \frac{e^t t}{2t} = \lim_{t\to 0} \frac{e^t t}{2t} = \lim_{t\to 0} \frac{e^t}{2} = \frac{1}{2} = \frac{1}{2}.$$

So, $E[X] = \frac{1}{2}$.

The second derivative of $M(t)$ with respect to t is:

$$M''(t) = \frac{d}{dt}\left(\frac{e^t(t-1) + 1}{t^2}\right).$$

DRAFT

After some calculation (similar to the first derivative), we find:

$$M''(t) = \frac{e^t(t^2 - 2t + 2) - 2(t-1)e^t}{t^3}.$$

Evaluating at $t = 0$:

$$M''(0) = \lim_{t\to 0} \frac{e^t(t^2 - 2t + 2) - 2(t-1)e^t}{t^3} = \frac{1}{3}.$$

So, $E[X^2] = \frac{1}{3}$.

Thus, the first two moments for a uniform random variable X on $[0, 1]$ are:

$$E[X] = \frac{1}{2} \text{ and } E[X^2] = \frac{1}{3}.$$

5.7.3 Probability Generating Function (PGF)

A Probability Generating Function (PGF) is a related concept to the Moment Generating Function (MGF), specifically designed for discrete random variables. The PGF provides a way to encode the probability distribution of a discrete random variable into a generating function. It is particularly useful for random variables that take non-negative integer values, such as the number of successes in a binomial distribution, or the number of events in a Poisson distribution.

Let X be a discrete random variable with probability mass function $p(x) = P(X = x)$ for $x = 0, 1, 2, \ldots$. The PGF of X, denoted by $G(s)$, is defined as:

$$G(s) = E[s^X] = \sum_{x=0}^{\infty} s^x p(x),$$

where s is a real or complex number for which the series converges.

Properties of the PGF

Let X be a discrete random variable taking non-negative integer values, and let $G(s)$ be its probability generating function, defined as:

$$G(s) = E[s^X] = \sum_{x=0}^{\infty} s^x P(x)$$

1. Normalization Property

The PGF at $s = 1$ is always equal to 1:

$$G(1) = \sum_{x=0}^{\infty} 1^x P(x) = \sum_{x=0}^{\infty} P(x) = 1$$

This holds for any probability distribution.

DRAFT

2. Probability Recovery

The probability that the random variable X takes the value x can be recovered by differentiating the PGF:

$$P(X = x) = \frac{1}{x!} \frac{d^x}{ds^x} G(s) \Big|_{s=0}$$

This formula allows for the extraction of individual probabilities from the PGF.

3. Expected Value (Mean)

The expected value $E[X]$ of the random variable X can be obtained by differentiating the PGF and evaluating at $s = 1$:

$$E[X] = \frac{d}{ds} G(s) \Big|_{s=1}$$

This gives the mean of the distribution directly from the PGF.

4. Variance

The variance $\text{Var}(X)$ can be derived using the PGF. First, compute the first and second derivatives of the PGF:

$$E[X^2] = \frac{d^2}{ds^2} G(s) \Big|_{s=1}$$

Then, the variance is:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

5. Sum of Independent Random Variables

If X_1 and X_2 are independent random variables, the PGF of their sum is the product of their individual PGFs:

$$G_{X_1+X_2}(s) = G_{X_1}(s) \cdot G_{X_2}(s)$$

This property is useful for dealing with sums of independent random variables.

6. PGF of a Constant

If X is a constant random variable, i.e., $P(X = c) = 1$ for some constant c, the PGF is:

$$G(s) = s^c$$

This reflects that the random variable always takes the value c, so the PGF has only one non-zero term at $x = c$.

DRAFT

7. Derivative Relations

The r-th moment $E[X^r]$ can be derived from the PGF by differentiating it r times and evaluating at $s = 1$:

$$E[X^r] = \frac{d^r}{ds^r} G(s)\bigg|_{s=1}$$

This formula is useful for calculating higher-order moments of the distribution.

Problem 5.21. *Consider a random variable X with parameter λ. The probability mass function of X is:*

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 2, 3, \ldots.$$

Find PGF, and hence mean and variance.

Solution

The PGF $G(s)$ is defined as:

$$G(s) = E[s^X] = \sum_{x=0}^{\infty} p(x) s^x.$$

Substituting the PMF $p(x)$:

$$G(s) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} s^x.$$

We can factor out $e^{-\lambda}$ as it is constant with respect to the sum:

$$G(s) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!}.$$

Recognize that the sum is the Taylor series expansion of $e^{\lambda s}$:

$$\sum_{x=0}^{\infty} \frac{(\lambda s)^x}{x!} = e^{\lambda s}.$$

Therefore, the PGF is:

$$G(s) = e^{-\lambda} \cdot e^{\lambda s} = e^{\lambda(s-1)}.$$

DRAFT

Finding the Mean

The mean $E[X]$ can be obtained by differentiating the PGF and evaluating at $s = 1$:

$$E[X] = G'(1).$$

First, compute the derivative of $G(s)$:

$$G(s) = e^{\lambda(s-1)}.$$

$$G'(s) = \lambda e^{\lambda(s-1)}.$$

Evaluating at $s = 1$:

$$G'(1) = \lambda e^{\lambda(1-1)} = \lambda e^0 = \lambda.$$

Thus, the mean $E[X]$ is λ.

Finding the Variance

The variance can be found using:

$$\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2.$$

Compute the second derivative of $G(s)$:

$$G''(s) = \lambda^2 e^{\lambda(s-1)}.$$

Evaluating at $s = 1$:

$$G''(1) = \lambda^2 e^{\lambda(1-1)} = \lambda^2 e^0 = \lambda^2.$$

Now, calculate the variance:

$$\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2.$$

Substitute the values:

$$\text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Thus, the variance $\text{Var}(X)$ is λ.

DRAFT

Applications

PGFs are used in various fields including:

- **Queueing Theory**: To analyze the number of customers in a queue.
- **Reliability Engineering**: To model system lifetimes.
- **Genetics**: To study inheritance patterns.

The PGF is a compact and powerful tool for handling problems involving sums of random variables and their distributions.

5.7.4 Characteristic Function (CF)

The characteristic function (CF) of a random variable X is a fundamental tool in probability theory, and it is closely related to the moment generating function. The characteristic function provides an alternative way to describe the distribution of X, and it is particularly useful in the study of sums of independent random variables. It is important in data science, particularly in areas related to probability theory, statistical inference, and stochastic processes.

The characteristic function $\varphi(t)$ of a random variable X is defined as:

$$\varphi(t) = E\left[e^{itX}\right],$$

where i is the imaginary unit ($i^2 = -1$) and t is a real number.

5.7.5 Key Properties of Characteristic Functions

Let X be a random variable with characteristic function $\phi(t)$.

1. **Existence:** The characteristic function always exists and is well-defined for all real t.
2. **Normalization:**
$$\phi(0) = E[e^{i\cdot 0\cdot X}] = E[1] = 1.$$
3. **Uniqueness:** The characteristic function uniquely determines the distribution of X. If two random variables have the same characteristic function, they have the same distribution.
4. **Addition of Independent Random Variables:** If X and Y are independent, then:
$$\phi_{X+Y}(t) = \phi(t) \cdot \phi_Y(t).$$

DRAFT

5. **Moment Generating Function Relationship:** The moment generating function $M(t)$ is related to the characteristic function by:

$$M(t) = E[e^{tX}] = \phi(-it).$$

6. **Inverse Relationship:** The distribution function $F(x)$ can be recovered from the characteristic function using the inverse Fourier transform:

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t) e^{-itx}\, dt.$$

7. **Derivatives and Moments:** The n-th moment of X, if it exists, is given by:

$$E[X^n] = i^{-n}\, \frac{d^n \phi(t)}{dt^n}\Big|_{t=0}.$$

The Characteristic Function (CF) and the Moment Generating Function (MGF) are both tools used in probability theory to describe the distribution of random variables. While they share some similarities, they also have key differences. A detailed comparisons are presented in Table 5.6.

Table 5.6: Comparison between MGF and CF

Feature	**MGF** $M(t) = E[e^{tX}]$	**CF** $\varphi(t) = E[e^{itX}]$
Existence	Not guaranteed (finite only if $E[e^{tX}]$ exists for all t)	Always exists (since $\lvert e^{itX}\rvert = 1$)
Range of t	Real t	Real t, but involves imaginary unit i
Moments	Differentiation gives moments directly	Differentiation gives moments with i adjustment
Uniqueness	Uniquely determines distribution if it exists	Uniquely determines distribution
Fourier Transform	No direct connection	Essentially the Fourier transform of the PDF
Application	Used for finding moments and cumulants, proving Central Limit Theorem	Used in distributional analysis, sums of random variables, proving Central Limit Theorem

Problem 5.22 (**Discrete Random Variable**). *Consider the Problem 5.19, the PMF of discrete random variable X is*

- $P(X = 0) = p$

DRAFT

- $P(X = 1) = 1 - p$

where $0 \leq p \leq 1$. *Find the characteristic function of* X, *and hence find mean and variance.*

Solution

1. Characteristic Function

The characteristic function $\varphi(t)$ of a discrete random variable X is defined as:

$$\varphi(t) = E[e^{itX}]$$

where E denotes the expectation and i is the imaginary unit.

For the given PMF:

$$\varphi(t) = E[e^{itX}] = \sum_x e^{itx} P(X = x)$$

Substituting the values for X:

$$\varphi(t) = e^{it \cdot 0} \cdot P(X = 0) + e^{it \cdot 1} \cdot P(X = 1)$$

$$\varphi(t) = e^0 \cdot p + e^{it} \cdot (1 - p)$$

$$\varphi(t) = p + (1 - p)e^{it}$$

2. Mean of X

The mean $E[X]$ can be derived from the characteristic function as follows:

$$E[X] = i \left. \frac{d}{dt} \varphi(t) \right|_{t=0}$$

First, compute the derivative of $\varphi(t)$:

$$\frac{d}{dt} \varphi(t) = \frac{d}{dt} \left(p + (1 - p)e^{it} \right)$$

$$\frac{d}{dt} \varphi(t) = (1 - p) \cdot ie^{it}$$

Evaluate at $t = 0$:

$$\left. \frac{d}{dt} \varphi(t) \right|_{t=0} = (1 - p) \cdot ie^{i \cdot 0} = (1 - p) \cdot i$$

$$E[X] = i \cdot (1 - p) \cdot i = 1 - p$$

DRAFT

3. Variance of X

To find the variance, we first need the second moment $E[X^2]$.

The second moment $E[X^2]$ can be derived from the characteristic function as follows:

$$E[X^2] = -\left.\frac{d^2}{dt^2}\varphi(t)\right|_{t=0}$$

Compute the second derivative of $\varphi(t)$:

$$\frac{d^2}{dt^2}\varphi(t) = \frac{d}{dt}\left((1-p)\cdot ie^{it}\right)$$

$$\frac{d^2}{dt^2}\varphi(t) = (1-p)\cdot i^2 e^{it} = -(1-p)e^{it}$$

Evaluate at $t = 0$:

$$\left.\frac{d^2}{dt^2}\varphi(t)\right|_{t=0} = -(1-p)e^{i\cdot 0} = -(1-p)$$

$$E[X^2] = -\left(-(1-p)\right) = 1-p$$

The variance $\mathrm{Var}(X)$ is given by:

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - (E[X])^2\\ &= (1-p) - (1-p)^2\\ &= (1-p) - (1-2p+p^2)\\ &= 1-p-1+2p-p^2\\ &= p-p^2\\ &= p(1-p)\end{aligned}$$

Hence,

- Characteristic function: $\varphi(t) = p + (1-p)e^{it}$
- Mean: $E[X] = 1-p$
- Variance: $\mathrm{Var}(X) = p(1-p)$

Problem 5.23 (**Continuous Random Variable**). *Consider the Problem 5.20, the PDF of X is given by:*

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1,\\ 0 & \text{otherwise.}\end{cases}$$

Find the characteristic function of X, and hence find the mean and variance.

DRAFT

Solution

Let's find the characteristic function of a uniform random variable X on the interval $[0, 1]$.

The characteristic function is:

$$\varphi(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx,$$

where $f(x)$ is the PDF of X.

For the uniform distribution on $[0, 1]$, the PDF is:

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the characteristic function becomes:

$$\varphi(t) = \int_0^1 e^{itx}\, dx.$$

The integral is straightforward to evaluate:

$$\varphi(t) = \int_0^1 e^{itx}\, dx = \left[\frac{e^{itx}}{it}\right]_0^1 = \frac{e^{it} - 1}{it}.$$

The characteristic function can be simplified as:

$$\varphi(t) = \frac{e^{it} - 1}{it}.$$

This can also be written as:

$$\varphi(t) = \frac{1}{t}\left(\frac{\sin(t)}{t} + i\frac{\cos(t) - 1}{t}\right), \quad \text{for } t \neq 0.$$

For $t = 0$, $\varphi(0) = 1$, which is consistent since the characteristic function always equals 1 at $t = 0$.

Mean

The mean $E[X]$ is:

$$E[X] = i\, \frac{d}{dt}\varphi(t)\bigg|_{t=0}$$

Compute the derivative of $\varphi(t)$:

$$\frac{d}{dt}\varphi(t) = \frac{d}{dt}\left(\frac{e^{it} - 1}{it}\right)$$

DRAFT

$$\frac{d}{dt}\varphi(t) = \frac{(it \cdot ie^{it}) - (e^{it} - 1)}{(it)^2}$$

$$\frac{d}{dt}\varphi(t) = \frac{-te^{it} + 1 - e^{it}}{-t^2}$$

$$\frac{d}{dt}\varphi(t) = \frac{1 - e^{it}(t+1)}{t^2}$$

Evaluate at $t = 0$:

$$\left.\frac{d}{dt}\varphi(t)\right|_{t=0} = \frac{1-1}{0^2} = 0$$

$$E[X] = 0$$

Variance

To find the variance, we first need the second moment $E[X^2]$:

$$E[X^2] = -\left.\frac{d^2}{dt^2}\varphi(t)\right|_{t=0}$$

Compute the second derivative:

$$\frac{d^2}{dt^2}\varphi(t) = \frac{d}{dt}\left(\frac{1 - e^{it}(t+1)}{t^2}\right)$$

$$\frac{d^2}{dt^2}\varphi(t) = \frac{t^2 \cdot (ie^{it}(t+1)) - (1 - e^{it}(t+1)) \cdot 2t}{t^4}$$

Evaluate at $t = 0$:

$$\left.\frac{d^2}{dt^2}\varphi(t)\right|_{t=0} = 2$$

$$E[X^2] = 2$$

The variance is:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

$$\mathrm{Var}(X) = 2 - 0^2$$

$$\mathrm{Var}(X) = 2$$

Hence,

- Characteristic function: $\varphi(t) = \frac{e^{it}-1}{it}$
- Mean: $E[X] = 0$
- Variance: $\mathrm{Var}(X) = 2$

DRAFT

5.7.6 Exercises

1. Consider a discrete random variable Y with the following probability mass function:
$$P(Y = k) = \begin{cases} \frac{1}{2} & \text{for } k = 1, \\ \frac{1}{2} & \text{for } k = 2. \end{cases}$$
   (a) Find the moment generating function $M_Y(t)$ of Y.
   (b) Using the moment generating function, find the mean and variance of Y.

2. Let Z be a continuous random variable with the probability density function:
$$f_Z(z) = \begin{cases} \frac{2z}{\theta^2} & \text{if } 0 \leq z \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$
   where $\theta > 0$ is a parameter.
   (a) Find the moment generating function $M_Z(t)$ of Z.
   (b) Find the characteristic function $\varphi_Z(t)$ of Z.
   (c) Using the moment generating function, determine the mean and variance of Z.

3. Let X be a random variable with parameters n and p, where n is the number of trials and p is the probability of success in each trial. The probability mass function of X is:
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$
   (a) Find the moment generating function $M(t)$ of X.
   (b) Using the moment generating function, find the mean and variance of X.

4. Consider a discrete random variable X that takes non-negative integer values with the following probability mass function:
$$P(X = k) = \frac{3^k}{\sum_{i=0}^{\infty} 3^i} \text{ for } k = 0, 1, 2, \ldots$$
   (a) Find the probability generating function (PGF) $G(s)$ of the random variable X.
   (b) Use the PGF to determine $E[X]$ and $\text{Var}(X)$.
   (c) Verify your results using the properties of the PGF.

DRAFT

5. Consider a continuous random variable W uniformly distributed over the interval $[0, a]$. The probability density function of W is:

$$f_W(w) = \begin{cases} \frac{1}{a} & \text{if } 0 \leq w \leq a, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find the characteristic function $\varphi_W(t)$ of W.

   (b) Find the moment generating function $M_W(t)$ of W.

   (c) Use the moment generating function to find the mean and variance of W.

6. Let V be a random variable with rate parameter λ. The probability density function is:

$$f_V(v) = \begin{cases} \lambda e^{-\theta v} & \text{if } v \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find the moment generating function $M_V(t)$ of V.

   (b) Find the characteristic function $\varphi_V(t)$ of V.

   (c) Using the moment generating function, determine the mean and variance of V.

5.8 Jointly Distributed Random Variables

Jointly distributed random variables are crucial in data science as they allow for the modeling and analysis of relationships between multiple variables simultaneously. Understanding these relationships is essential for predicting outcomes, identifying correlations, and constructing probabilistic models that capture real-world complexities. Joint distributions enable informed decisions based on the combined behavior of multiple variables, which is vital for developing accurate and robust predictive models.

In probability theory, two or more random variables are jointly distributed if there is a joint probability distribution describing their behavior. For two random variables X and Y, the joint probability distribution provides the probability that X takes a specific value x and Y takes a specific value y simultaneously.

5.8.1 Joint Probability Mass Function (PMF)

For discrete random variables X and Y, the joint probability mass function $p_{X,Y}(x, y)$ is defined as:

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

DRAFT

The joint probability mass function must satisfy the condition:

$$\sum_x \sum_y p_{X,Y}(x,y) = 1.$$

The joint cumulative distribution function is defined as:

$$F(x,y) = P(X \leq x, Y \leq y).$$

For discrete random variables:

$$F(x,y) = \sum_{X \leq x} \sum_{Y \leq y} p_{X,Y}(x,y).$$

If $p_{X,Y}(i,j) = P(X = i, Y = j)$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, then

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{i=1}^{x} \sum_{j=1}^{y} p_{X,Y}(i,j)$$

5.8.2 Example: Daily Customers and Product Sales

A small retail store is analyzing the relationship between the number of customers visiting the store and the number of products sold during a day. Understanding this relationship helps the store owner make better decisions about inventory management and promotional strategies.

Let the random variable X denote the **Number of Customers per Day**, taking values 0, 1, and 2 (for simplicity, we consider a scenario where the store experiences low traffic during specific time periods). Let the random variable Y represent the **Number of Products Sold per Day**, also taking values 0, 1, and 2. These two random variables are considered jointly distributed.

The **joint probability mass function** $p_{X,Y}(x,y)$ for these variables, derived from historical sales data, is given in the table below:

Table 5.7: Joint Probability Mass Function $p_{X,Y}(x,y)$ for Customers (X) and Products Sold (Y)

		Products Sold (Y)		
		0	**1**	**2**
Customers per Day (X)	**0**	0.10	0.00	0.00
	1	0.15	0.20	0.10
	2	0.05	0.15	0.25

DRAFT

Verification of Joint PMF

We first verify that the joint probability mass function satisfies the required condition:

$$\begin{aligned}\sum_x \sum_y p_{X,Y}(x,y) &= 0.10 + 0.00 + 0.00 + 0.15 + 0.20 + 0.10 \\ &\quad + 0.05 + 0.15 + 0.25 \\ &= 1.00.\end{aligned}$$

This confirms that $p_{X,Y}(x,y)$ is a valid joint probability mass function.

Joint Cumulative Distribution Function

The joint cumulative distribution function $F(x,y) = P(X \leq x, Y \leq y)$ can be computed for all values of x and y. We calculate selected values below:

Example 1: Calculate $F(1,1) = P(X \leq 1, Y \leq 1)$.

$$\begin{aligned}F(1,1) &= \sum_{i=0}^{1} \sum_{j=0}^{1} p_{X,Y}(i,j) \\ &= p_{X,Y}(0,0) + p_{X,Y}(0,1) + p_{X,Y}(1,0) + p_{X,Y}(1,1) \\ &= 0.10 + 0.00 + 0.15 + 0.20 \\ &= 0.45.\end{aligned}$$

Example 2: Calculate $F(2,1) = P(X \leq 2, Y \leq 1)$.

$$\begin{aligned}F(2,1) &= \sum_{i=0}^{2} \sum_{j=0}^{1} p_{X,Y}(i,j) \\ &= p_{X,Y}(0,0) + p_{X,Y}(0,1) + p_{X,Y}(1,0) + p_{X,Y}(1,1) \\ &\quad + p_{X,Y}(2,0) + p_{X,Y}(2,1) \\ &= 0.10 + 0.00 + 0.15 + 0.20 + 0.05 + 0.15 \\ &= 0.65.\end{aligned}$$

Example 3: Calculate $F(2,2) = P(X \leq 2, Y \leq 2)$.

$$\begin{aligned}F(2,2) &= \sum_{i=0}^{2} \sum_{j=0}^{2} p_{X,Y}(i,j) \\ &= \sum_x \sum_y p_{X,Y}(x,y) \\ &= 1.00.\end{aligned}$$

As expected, $F(2,2) = 1$ since this represents the probability of the entire sample space.

DRAFT

5.8.3 Joint Probability Density Function

For continuous random variables X and Y, the joint probability density function $f_{X,Y}(x, y)$ is defined as:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y).$$

The joint probability density function must satisfy the condition:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

The probability that $a \leq X \leq b$ and $c \leq Y \leq d$ is obtained from the joint probability density function as:

$$\int_{x=a}^{b} \int_{y=c}^{d} f(x, y)\, dy\, dx$$

For continuous random variables:

$$F(x, y) = \int_{w=-\infty}^{x} \int_{z=-\infty}^{y} f(w, z)\, dz\, dw$$

5.8.4 Example: Mineral Deposits

To evaluate the economic feasibility of mining in a specific region, a mining company collects ore samples from the site and measures their zinc and iron content. Let the random variable X represent the zinc content, ranging from 0.5 to 1.5, and the random variable Y represent the iron content, ranging from 20.0 to 35.0. Suppose the joint probability density function of X and Y is given by

$$f(x, y) = \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10,000}$$

for $0.5 \leq x \leq 1.5$ and $20.0 \leq y \leq 35.0$.

To verify the validity of this joint probability density function, we need to ensure that $f(x, y) \geq 0$ within the defined state space and that

$$\int_{0.5}^{1.5} \int_{20.0}^{35.0} f(x, y)\, dy\, dx = 1.$$

This joint probability density function provides comprehensive information about the joint probabilistic behavior of the random variables X and Y. For instance, the probability that a randomly selected ore sample has a zinc content between 0.8 and 1.0 and an iron content between 25 and 30 is given by

DRAFT

$$\int_{0.8}^{1.0} \int_{25.0}^{30.0} f(x, y)\, dy\, dx,$$

which evaluates to 0.092. Thus, only about 9% of the ore at this location has mineral levels within these specified ranges.

5.8.5 Marginal Distributions

The marginal distributions of X and Y provide the probability distribution for each random variable independently, by summing or integrating over the values of the other variable.

For discrete random variables:

$$p_X(x) = \sum_y p(x, y)$$

$$p_Y(y) = \sum_x p(x, y)$$

For continuous random variables:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

Using these marginal distributions, we can easily find the mean, variance, and standard deviation for X and Y individually.

5.8.6 Example: Daily Customers and Product Sales

Refer to the example given in 5.8.2. To find the marginal distributions of X (Number of Customers per Day) and Y (Number of Products Sold per Day) for the example **Daily Customers and Product Sales**, as discussed in Section 5.8.2, we need to sum the probabilities across rows for X and across columns for Y.

The joint probability mass function with its marginal distributions is shown in the table below:

DRAFT

Table 5.8: Joint Probability Mass Function $p(x, y)$ for Customers (X) and Products Sold (Y) with Marginal Distributions

		Products Sold (Y)			
		0	**1**	**2**	$p_X(x)$
Customers per Day (X)	**0**	0.10	0.00	0.00	**0.10**
	1	0.15	0.20	0.10	**0.45**
	2	0.05	0.15	0.25	**0.45**
	$p_Y(y)$	**0.30**	**0.35**	**0.35**	**1.00**

Marginal PMF for X (Number of Customers per Day):

$$p_X(0) = p(0,0) + p(0,1) + p(0,2) = 0.10 + 0.00 + 0.00 = 0.10$$
$$p_X(1) = p(1,0) + p(1,1) + p(1,2) = 0.15 + 0.20 + 0.10 = 0.45$$
$$p_X(2) = p(2,0) + p(2,1) + p(2,2) = 0.05 + 0.15 + 0.25 = 0.45$$

Marginal PMF for Y (Number of Products Sold per Day):

$$p_Y(0) = p(0,0) + p(1,0) + p(2,0) = 0.10 + 0.15 + 0.05 = 0.30$$
$$p_Y(1) = p(0,1) + p(1,1) + p(2,1) = 0.00 + 0.20 + 0.15 = 0.35$$
$$p_Y(2) = p(0,2) + p(1,2) + p(2,2) = 0.00 + 0.10 + 0.25 = 0.35$$

Mean of X

The mean (expected value) of X is calculated using its marginal PMF $p_X(x)$:

$$\begin{aligned}\mu_X &= \sum_x x \cdot p_X(x) \\ &= (0 \cdot 0.10) + (1 \cdot 0.45) + (2 \cdot 0.45) \\ &= 0 + 0.45 + 0.90 \\ &= 1.35\end{aligned}$$

On average, there are 1.35 customers per day.

DRAFT

Expected Value of X^2

To calculate the variance, we first find $E(X^2)$:

$$\begin{aligned} E(X^2) &= \sum_x x^2 \cdot p_X(x) \\ &= (0^2 \cdot 0.10) + (1^2 \cdot 0.45) + (2^2 \cdot 0.45) \\ &= (0 \cdot 0.10) + (1 \cdot 0.45) + (4 \cdot 0.45) \\ &= 0 + 0.45 + 1.80 \\ &= 2.25 \end{aligned}$$

Variance of X

$$\text{Var}(X) = E(X^2) - (\mu_X)^2 = 2.25 - (1.35)^2 = 2.25 - 1.8225 = 0.4275$$

Standard Deviation of X

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{0.4275} \approx 0.6538$$

Mean of Y

The mean (expected value) of Y is calculated using its marginal PMF $p_Y(y)$:

$$\begin{aligned} \mu_Y &= \sum_y y \cdot p_Y(y) \\ &= (0 \cdot 0.30) + (1 \cdot 0.35) + (2 \cdot 0.35) \\ &= 0 + 0.35 + 0.70 \\ &= 1.05 \end{aligned}$$

On average, 1.05 products are sold per day.

Expected Value of Y^2

To calculate the variance, we first find $E(Y^2)$:

$$\begin{aligned} E(Y^2) &= \sum_y y^2 \cdot p_Y(y) \\ &= (0^2 \cdot 0.30) + (1^2 \cdot 0.35) + (2^2 \cdot 0.35) \\ &= (0 \cdot 0.30) + (1 \cdot 0.35) + (4 \cdot 0.35) \\ &= 0 + 0.35 + 1.40 \\ &= 1.75 \end{aligned}$$

Variance of Y

$$\text{Var}(Y) = E(Y^2) - (\mu_Y)^2 = 1.75 - (1.05)^2 = 1.75 - 1.1025 = 0.6475$$

DRAFT

Standard Deviation of Y

$$\sigma_Y = \sqrt{\text{Var}(Y)} = \sqrt{0.6475} \approx 0.8047$$

5.8.7 Example: Daily Customers and Product Sales - Conditional Distributions

Refer to the example given in 5.8.2. To find the conditional mean and variance of X given Y and Y given X, we use the definitions of conditional expectations and variances. The conditional probability mass function $p_{X|Y}(x|y)$ is given by $p(x, y)/p_Y(y)$, and $p_{Y|X}(y|x)$ is given by $p(x, y)/p_X(x)$.

The joint PMF and marginal PMFs are repeated here for convenience:

Table 5.9: Joint Probability Mass Function $p(x, y)$ for Customers (X) and Products Sold (Y) with Marginal Distributions

		Products Sold (Y)			
		0	**1**	**2**	$p_X(x)$
Customers per Day (X)	**0**	0.10	0.00	0.00	**0.10**
	1	0.15	0.20	0.10	**0.45**
	2	0.05	0.15	0.25	**0.45**
	$p_Y(y)$	**0.30**	**0.35**	**0.35**	**1.00**

Conditional Distribution of X given Y

Case 1: $Y = 0$

Given $Y = 0$, the marginal probability is $p_Y(0) = 0.30$. The conditional PMF $p_{X|Y}(x|0)$ is:

$$p_{X|Y}(0|0) = \frac{p(0,0)}{p_Y(0)} = \frac{0.10}{0.30} \approx 0.3333$$
$$p_{X|Y}(1|0) = \frac{p(1,0)}{p_Y(0)} = \frac{0.15}{0.30} = 0.5000$$
$$p_{X|Y}(2|0) = \frac{p(2,0)}{p_Y(0)} = \frac{0.05}{0.30} \approx 0.1667$$

The conditional distribution of X given $Y = 0$:

DRAFT

Table 5.10: The conditional distribution of X given $Y = 0$

x	0	1	2
$p_{X\|Y}(x\|0)$	0.3333	0.5000	0.1667

For $Y = 0$:

$$\begin{aligned} E(X|Y=0) &= (0 \cdot 0.3333) + (1 \cdot 0.5000) + (2 \cdot 0.1667) \\ &= 0 + 0.5000 + 0.3334 = 0.8334 \end{aligned}$$

$$\begin{aligned} E(X^2|Y=0) &= (0^2 \cdot 0.3333) + (1^2 \cdot 0.5000) + (2^2 \cdot 0.1667) \\ &= (0 \cdot 0.3333) + (1 \cdot 0.5000) + (4 \cdot 0.1667) \\ &= 0 + 0.5000 + 0.6668 = 1.1668 \end{aligned}$$

$$\begin{aligned} \text{Var}(X|Y=0) &= E(X^2|Y=0) - (E(X|Y=0))^2 \\ &= 1.1668 - (0.8334)^2 \approx 1.1668 - 0.6946 = 0.4722 \end{aligned}$$

Hence, the standard deviation of X given $Y = 0$ is $\sqrt{0.4722} \approx 0.6872$.

Case 2: $Y = 1$

Given $Y = 1$, the marginal probability is $p_Y(1) = 0.35$. The conditional PMF $p_{X|Y}(x|1)$ is:

$$\begin{aligned} p_{X|Y}(0|1) &= \frac{p(0,1)}{p_Y(1)} = \frac{0.00}{0.35} = 0.00 \\ p_{X|Y}(1|1) &= \frac{p(1,1)}{p_Y(1)} = \frac{0.20}{0.35} \approx 0.5714 \\ p_{X|Y}(2|1) &= \frac{p(2,1)}{p_Y(1)} = \frac{0.15}{0.35} \approx 0.4286 \end{aligned}$$

The conditional distribution of X given $Y = 1$:

Table 5.11: The conditional distribution of X given $Y = 1$

x	0	1	2
$p_{X\|Y}(x\|1)$	0.00	0.5714	0.4286

For $Y = 1$:

$$\begin{aligned} E(X|Y=1) &= (0 \cdot 0.00) + (1 \cdot 0.5714) + (2 \cdot 0.4286) \\ &= 0 + 0.5714 + 0.8572 = 1.4286 \end{aligned}$$

DRAFT

$$\begin{aligned} E(X^2|Y=1) &= (0^2 \cdot 0.00) + (1^2 \cdot 0.5714) + (2^2 \cdot 0.4286) \\ &= (0 \cdot 0.00) + (1 \cdot 0.5714) + (4 \cdot 0.4286) \\ &= 0 + 0.5714 + 1.7144 = 2.2858 \end{aligned}$$

$$\begin{aligned} \text{Var}(X|Y=1) &= E(X^2|Y=1) - (E(X|Y=1))^2 \\ &= 2.2858 - (1.4286)^2 \approx 2.2858 - 2.0408 \\ &= 0.2450 \end{aligned}$$

Hence, the standard deviation of X given $Y = 1$ is $\sqrt{0.2450} \approx 0.4950$.

Case 3: $Y = 2$

Given $Y = 2$, the marginal probability is $p_Y(2) = 0.35$. The conditional PMF $p_{X|Y}(x|2)$ is:

$$\begin{aligned} p_{X|Y}(0|2) &= \frac{p(0,2)}{p_Y(2)} = \frac{0.00}{0.35} = 0.00 \\ p_{X|Y}(1|2) &= \frac{p(1,2)}{p_Y(2)} = \frac{0.10}{0.35} \approx 0.2857 \\ p_{X|Y}(2|2) &= \frac{p(2,2)}{p_Y(2)} = \frac{0.25}{0.35} \approx 0.7143 \end{aligned}$$

The conditional distribution of X given $Y = 2$:

Table 5.12: The conditional distribution of X given $Y = 2$

x	0	1	2
$p_{X\|Y}(x\|2)$	0.00	0.2857	0.7143

For $Y = 2$:

$$\begin{aligned} E(X|Y=2) &= (0 \cdot 0.00) + (1 \cdot 0.2857) + (2 \cdot 0.7143) \\ &= 0 + 0.2857 + 1.4286 = 1.7143 \end{aligned}$$

$$\begin{aligned} E(X^2|Y=2) &= (0^2 \cdot 0.00) + (1^2 \cdot 0.2857) + (2^2 \cdot 0.7143) \\ &= (0 \cdot 0.00) + (1 \cdot 0.2857) + (4 \cdot 0.7143) \\ &= 0 + 0.2857 + 2.8572 = 3.1429 \end{aligned}$$

$$\begin{aligned} \text{Var}(X|Y=2) &= E(X^2|Y=2) - (E(X|Y=2))^2 \\ &= 3.1429 - (1.7143)^2 \approx 3.1429 - 2.9388 \\ &= 0.2041 \end{aligned}$$

Hence, the standard deviation of X given $Y = 2$ is $\sqrt{0.2041} \approx 0.4518$.

DRAFT

Conditional Distribution of Y given X

Case 1: $X = 0$

Given $X = 0$, the marginal probability is $p_X(0) = 0.10$. The conditional PMF $p_{Y|X}(y|0)$ is:

$$p_{Y|X}(0|0) = \frac{p(0,0)}{p_X(0)} = \frac{0.10}{0.10} = 1.00$$
$$p_{Y|X}(1|0) = \frac{p(0,1)}{p_X(0)} = \frac{0.00}{0.10} = 0.00$$
$$p_{Y|X}(2|0) = \frac{p(0,2)}{p_X(0)} = \frac{0.00}{0.10} = 0.00$$

The conditional distribution of Y given $X = 0$:

Suppose that the

Table 5.13: The conditional distribution of Y given $X = 0$

y	0	1	2
$p_{Y\|X}(y\|0)$	1.00	0.00	0.00

For $X = 0$:

$$E(Y|X = 0) = (0 \cdot 1.00) + (1 \cdot 0.00) + (2 \cdot 0.00)$$
$$= 0 + 0 + 0 = 0.00$$

$$E(Y^2|X = 0) = (0^2 \cdot 1.00) + (1^2 \cdot 0.00) + (2^2 \cdot 0.00)$$
$$= 0 + 0 + 0 = 0.00$$

$$\text{Var}(Y|X = 0) = E(Y^2|X = 0) - (E(Y|X = 0))^2 = 0.00 - (0.00)^2 = 0.00$$

Hence, the standard deviation of Y given $X = 0$ is $\sqrt{0.00} = 0.00$. This makes sense: if there are no customers, no products can be sold.

Case 2: $X = 1$

Given $X = 1$, the marginal probability is $p_X(1) = 0.45$. The conditional PMF $p_{Y|X}(y|1)$ is:

$$p_{Y|X}(0|1) = \frac{p(1,0)}{p_X(1)} = \frac{0.15}{0.45} \approx 0.3333$$
$$p_{Y|X}(1|1) = \frac{p(1,1)}{p_X(1)} = \frac{0.20}{0.45} \approx 0.4444$$
$$p_{Y|X}(2|1) = \frac{p(1,2)}{p_X(1)} = \frac{0.10}{0.45} \approx 0.2222$$

DRAFT

The conditional distribution of Y given $X = 1$:

Table 5.14: The conditional distribution of Y given $X = 1$

y	0	1	2
$p_{Y\|X}(y\|1)$	0.3333	0.4444	0.2222

For $X = 1$:

$$\begin{aligned} E(Y|X=1) &= (0 \cdot 0.3333) + (1 \cdot 0.4444) + (2 \cdot 0.2222) \\ &= 0 + 0.4444 + 0.4444 = 0.8888 \end{aligned}$$

$$\begin{aligned} E(Y^2|X=1) &= (0^2 \cdot 0.3333) + (1^2 \cdot 0.4444) + (2^2 \cdot 0.2222) \\ &= (0 \cdot 0.3333) + (1 \cdot 0.4444) + (4 \cdot 0.2222) \\ &= 0 + 0.4444 + 0.8888 = 1.3332 \end{aligned}$$

$$\begin{aligned} \text{Var}(Y|X=1) &= E(Y^2|X=1) - (E(Y|X=1))^2 \\ &= 1.3332 - (0.8888)^2 \approx 1.3332 - 0.7900 \\ &= 0.5432 \end{aligned}$$

Hence, the standard deviation of Y given $X = 1$ is $\sqrt{0.5432} \approx 0.7370$.

Case 3: $X = 2$

Given $X = 2$, the marginal probability is $p_X(2) = 0.45$. The conditional PMF $p_{Y|X}(y|2)$ is:

$$\begin{aligned} p_{Y|X}(0|2) &= \frac{p(2,0)}{p_X(2)} = \frac{0.05}{0.45} \approx 0.1111 \\ p_{Y|X}(1|2) &= \frac{p(2,1)}{p_X(2)} = \frac{0.15}{0.45} \approx 0.3333 \\ p_{Y|X}(2|2) &= \frac{p(2,2)}{p_X(2)} = \frac{0.25}{0.45} \approx 0.5556 \end{aligned}$$

The conditional distribution of Y given $X = 2$:

Table 5.15: The conditional distribution of Y given $X = 2$

y	0	1	2
$p_{Y\|X}(y\|2)$	0.1111	0.3333	0.5556

DRAFT

For $X = 2$:

$$\begin{aligned} E(Y|X=2) &= (0 \cdot 0.1111) + (1 \cdot 0.3333) + (2 \cdot 0.5556) \\ &= 0 + 0.3333 + 1.1112 = 1.4445 \end{aligned}$$

$$\begin{aligned} E(Y^2|X=2) &= (0^2 \cdot 0.1111) + (1^2 \cdot 0.3333) + (2^2 \cdot 0.5556) \\ &= (0 \cdot 0.1111) + (1 \cdot 0.3333) + (4 \cdot 0.5556) \\ &= 0 + 0.3333 + 2.2224 = 2.5557 \end{aligned}$$

$$\begin{aligned} \text{Var}(Y|X=2) &= E(Y^2|X=2) - (E(Y|X=2))^2 \\ &= 2.5557 - (1.4445)^2 \approx 2.5557 - 2.0866 \\ &= 0.4691 \end{aligned}$$

Hence, the standard deviation of Y given $X = 2$ is $\sqrt{0.4691} \approx 0.6849$.

5.8.8 Example: Mineral Deposits

Given a sample of ore with a zinc content of $X = 0.55$, what can be inferred about its iron content? The information regarding the iron content Y is encapsulated in the conditional probability density function, which is expressed as:

$$f_{Y|X=0.55}(y) = \frac{f(0.55, y)}{f_X(0.55)}$$

Here, the denominator represents the marginal distribution of the zinc content X evaluated at 0.55. Evaluating $f_X(0.55)$:

$$f_X(0.55) = \frac{57}{40} - \frac{51(0.55 - 1.00)^2}{10} = 0.39225$$

Thus, the conditional probability density function becomes:

$$f_{Y|X=0.55}(y) = \frac{f(0.55, y)}{0.39225} = \frac{\frac{39}{400} - \frac{17(0.55-1.00)^2}{50} - \frac{(y-25)^2}{10{,}000}}{0.39225}$$

Simplifying, we get:

$$f_{Y|X=0.55}(y) = 0.073 - \frac{(y-25)^2}{3922.5}$$

for $20.0 \leq y \leq 35.0$. It can be easily find the conditional expectation of the iron content, which is calculated to be 27.14, and the conditional standard deviation, which is 4.14.

DRAFT

5.8.9 Independent Random Variables

Just as two events A and B are considered independent if they are "unrelated" to each other, two random variables X and Y are deemed independent if the value taken by one random variable is "unrelated" to the value taken by the other. Specifically, in the context of data science, random variables are independent if the distribution of one random variable does not depend on the value taken by the other random variable.

Independent Random Variables

- For discrete random variables, independence means that the joint probability mass function can be expressed as the product of their individual PMF's:

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y).$$

- For continuous random variables, independence means that the joint probability density function can be expressed as the product of their individual PDF's:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

Example:

- Let X be the result of rolling a fair six-sided die, and Y be the result of flipping a fair coin, where X can take values 1 through 6, and Y can take values 0 (tails) and 1 (heads). The events are independent, so:

$$P(X = 3 \text{ and } Y = 1) = P(X = 3) \cdot P(Y = 1) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Problem 5.24. *It is known that the ratio of gallium to arsenide does not affect the functioning of gallium-arsenide wafers, which are the main components of microchips. Let X denote the ratio of gallium to arsenide and Y denote the functional wafers retrieved during a 1-hour period. X and Y are independent random variables with the joint density function*

$$f(x, y) = \begin{cases} \frac{x(1+3y^2)}{4}, & 0 < x < 2,\ 0 < y < 1, \\ 0, & \textit{elsewhere.} \end{cases}$$

Show that X and Y are independent random variables.

Solution

To show that X and Y are independent random variables, we need to verify that the joint density function $f(x, y)$ can be factored into the product of

DRAFT

the marginal density functions $f_X(x)$ and $f_Y(y)$. Specifically, X and Y are independent if and only if the joint density function $f(x, y)$ can be written as:

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

Marginal density function $f_X(x)$:

To find $f_X(x)$, integrate the joint density function $f(x, y)$ over the possible values of y:

$$f_X(x) = \int_0^1 f(x, y)\, dy.$$

Given the joint density function $f(x, y) = \frac{x(1+3y^2)}{4}$ for $0 < x < 2$ and $0 < y < 1$, compute:

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{x(1+3y^2)}{4}\, dy \\ &= \frac{x}{4} \int_0^1 (1+3y^2)\, dy \\ &= \frac{x}{4} \left(\int_0^1 1\, dy + \int_0^1 3y^2\, dy \right) \\ &= \frac{x}{4}(1+1) \\ &= \frac{x}{2}. \end{aligned}$$

So, the marginal density function for X is:

$$f_X(x) = \frac{x}{2}, \quad 0 < x < 2.$$

Marginal density function $f_Y(y)$:

To find $f_Y(y)$, integrate the joint density function $f(x, y)$ over the possible values of x:

$$\begin{aligned} f_Y(y) &= \int_0^2 f(x, y)\, dx \\ &= \int_0^2 \frac{x(1+3y^2)}{4}\, dx \\ &= \frac{1+3y^2}{4} \int_0^2 x\, dx \\ &= \frac{1+3y^2}{4} \cdot 2 = \frac{1+3y^2}{2}. \end{aligned}$$

DRAFT

Therefore, the marginal density function for Y is:

$$f_Y(y) = \frac{1+3y^2}{2}, \quad 0 < y < 1.$$

Verify independence

Check if $f(x, y)$ can be written as $f_X(x) \cdot f_Y(y)$:

$$f_X(x) \cdot f_Y(y) = \left(\frac{x}{2}\right) \cdot \left(\frac{1+3y^2}{2}\right) = \frac{x(1+3y^2)}{4}.$$

This matches the given joint density function $f(x, y)$.

Since $f(x, y) = f_X(x) \cdot f_Y(y)$, the random variables X and Y are independent.

□

5.8.10 Covariance and Correlation

Covariance

Covariance is essential for understanding and quantifying relationships between variables, which is a cornerstone of many data science techniques and analyses. It measures the joint variability of two random variables, indicating how they change together.

Covariance: For two random variables X and Y, the covariance is defined as:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

or equivalently, the computationally often more convenient form:

$$\text{Cov}(X, Y) = E[XY] - E(X)E(Y).$$

Interpretation of Covariance:

- A **positive covariance** indicates that X and Y tend to move in the same direction. When X is above its mean, Y tends to be above its mean, and vice versa.
- A **negative covariance** indicates that X and Y tend to move in opposite directions. When X is above its mean, Y tends to be below its mean, and vice versa.
- A **covariance close to zero** suggests little or no linear relationship between X and Y.

If X and Y are independent, then $\text{Cov}(X, Y) = 0$. However, a covariance of zero does not necessarily imply independence; it only implies an absence of a

DRAFT

linear relationship. Non-linear relationships might still exist. The magnitude of the covariance is difficult to interpret on its own because it is influenced by the scale of the variables. For instance, if X and Y are measured in different units (e.g., height in meters and weight in kilograms), their covariance will be affected by these units.

Correlation

Correlation is a normalized form of covariance that measures the strength and direction of the linear relationship between two random variables. This normalization makes correlation a more interpretable metric, as it is unitless and its value always falls within a specific range. It's widely used in statistical analysis, machine learning, and data visualization to reveal and quantify relationships that might otherwise be obscured by differences in scale or units.

Correlation: For two random variables X and Y, the correlation coefficient (often denoted as $\rho_{X,Y}$) is defined as:

$$\text{Corr}(X, Y) = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Interpretation of the Correlation Coefficient $\rho_{X,Y}$: The correlation coefficient $\rho_{X,Y}$ ranges from -1 to $+1$.

- A value of $+1$ implies a **perfect positive linear relationship**. This means that as X increases, Y increases proportionally in a perfectly predictable straight line.
- A value of -1 implies a **perfect negative linear relationship**. This means that as X increases, Y decreases proportionally in a perfectly predictable straight line.
- A value of 0 implies **no linear relationship**. It means there's no tendency for X and Y to increase or decrease together in a straight-line fashion. However, as with covariance, it does not rule out non-linear relationships.
- Values **between** 0 **and** $+1$ indicate a **positive linear relationship** of varying strength:
  - Close to 0: Weak positive linear relationship.
  - Close to $+1$: Strong positive linear relationship.
- Values **between** -1 **and** 0 indicate a **negative linear relationship** of varying strength:

DRAFT

- Close to 0: Weak negative linear relationship.
- Close to -1: Strong negative linear relationship.

It's crucial to remember that correlation measures only *linear* relationships. A high correlation does not imply causation, and a low correlation does not necessarily mean the variables are unrelated, only that they do not have a strong *linear* association.

Problem 5.25. *Consider the **Daily Customers and Product Sales** example, as introduced in Section 5.8.2, where the random variable X denotes the **Number of Customers per Day** (taking values 0, 1, and 2), and the random variable Y represents the **Number of Products Sold per Day** (also taking values 0, 1, and 2). The joint probability mass function $p(x, y)$ for these variables is given in the table below:*

Table 5.16: Joint Probability Mass Function $p_{X,Y}(x, y)$ for Customers (X) and Products Sold (Y)

		Products Sold (Y)		
		0	**1**	**2**
Customers per Day (X)	**0**	0.10	0.00	0.00
	1	0.15	0.20	0.10
	2	0.05	0.15	0.25

(a). Find the covariance of X and Y.

(b). Find the correlation of X and Y.

Solution

We recall the necessary marginal means and standard deviations calculated previously for this example:

- Mean of X, $\mu_X = E(X) = 1.35$
- Mean of Y, $\mu_Y = E(Y) = 1.05$
- Standard Deviation of X, $\sigma_X = \sqrt{\text{Var}(X)} \approx 0.6538$
- Standard Deviation of Y, $\sigma_Y = \sqrt{\text{Var}(Y)} \approx 0.8047$

DRAFT

(a) Covariance of X and Y:

The covariance of X and Y is defined as

$$\text{Cov}(X, Y) = E[XY] - E(X)E(Y).$$

We need to compute $E(XY)$ by summing the products of x, y, and their corresponding joint probabilities:

$$E(XY) = \sum_x \sum_y x \cdot y \cdot p(x, y)$$

Calculating each term:

$$\begin{aligned} E(XY) &= (0 \cdot 0 \cdot 0.10) + (0 \cdot 1 \cdot 0.00) + (0 \cdot 2 \cdot 0.00) \\ &\quad + (1 \cdot 0 \cdot 0.15) + (1 \cdot 1 \cdot 0.20) + (1 \cdot 2 \cdot 0.10) \\ &\quad + (2 \cdot 0 \cdot 0.05) + (2 \cdot 1 \cdot 0.15) + (2 \cdot 2 \cdot 0.25) \\ &= 0 + 0 + 0 \\ &\quad + 0 + 0.20 + 0.20 \\ &\quad + 0 + 0.30 + 1.00 \\ &= 1.70 \end{aligned}$$

Therefore, the expected value $E(XY)$ is 1.70. Now, we can calculate the covariance:

$$\text{Cov}(X, Y) = E[XY] - E(X)E(Y) = 1.70 - (1.35 \times 1.05)$$

$$\text{Cov}(X, Y) = 1.70 - 1.4175 = 0.2825$$

The positive value of 0.2825 indicates that there is a positive relationship between the Number of Customers per Day (X) and the Number of Products Sold per Day (Y). As the number of customers increases, the number of products sold tends to increase, and vice versa.

(b) Correlation of X and Y:

To get a more standardized measure of the relationship between X and Y, we compute the correlation coefficient:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Substitute the calculated covariance and the standard deviations:

$$\rho_{X,Y} = \frac{0.2825}{0.6538 \times 0.8047}$$

$$\rho_{X,Y} = \frac{0.2825}{0.52611286} \approx 0.5370$$

DRAFT

The correlation coefficient of approximately 0.5370 suggests a moderate positive linear relationship between the number of customers and the number of products sold. This means that days with more customers tend to be days with more products sold, which aligns with expectations for a retail store where customer traffic is positively associated with sales volume. □

Problem 5.26. *Consider two continuous random variables X and Y with the following joint probability density function:*

$$f_{X,Y}(x,y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Are X and Y independent?

(b) Find the covariance $Cov(X,Y)$.

(c) Find the correlation $Corr(X,Y)$.

Solution

(a). Are X and Y independent?

To check if X and Y are independent, we need to verify if the joint PDF factorizes into the product of the marginal PDFs of X and Y. That is, we need to check if:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

Marginal PDF of X:

The marginal PDF of X is obtained by integrating the joint PDF over all possible values of y:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)\, dy.$$

For $0 \le x \le 1$, we compute:

$$f_X(x) = \int_0^1 4xy\, dy = 4x \int_0^1 y\, dy = 4x \left[\frac{y^2}{2}\right]_0^1 = 4x \cdot \frac{1}{2} = 2x.$$

Thus, the marginal PDF of X is:

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

DRAFT

Marginal PDF of Y:

Similarly, the marginal PDF of Y is obtained by integrating the joint PDF over all possible values of x:

$$f_Y(y) = \int_0^1 f_{X,Y}(x, y)\, dx.$$

For $0 \leq y \leq 1$, we compute:

$$f_Y(y) = \int_0^1 4xy\, dx = 4y \int_0^1 x\, dx = 4y \left[\frac{x^2}{2}\right]_0^1 = 4y \cdot \frac{1}{2} = 2y.$$

Thus, the marginal PDF of Y is:

$$f_Y(y) = \begin{cases} 2y & \text{if } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Check for Independence:

Now we check if the joint PDF factorizes as the product of the marginal PDFs. We compute:

$$f_X(x) f_Y(y) = (2x)(2y) = 4xy.$$

Since $f_{X,Y}(x, y) = 4xy$ (which matches $f_X(x) f_Y(y)$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$), we conclude that X and Y are **independent**.

(b). Find the Covariance Cov(X, Y):

The covariance is defined as:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

Compute $E[X]$ and $E[Y]$:

First, we calculate the expected values $E[X]$ and $E[Y]$.

For X:

$$E[X] = \int_0^1 x f_X(x)\, dx = \int_0^1 x(2x)\, dx = 2 \int_0^1 x^2\, dx = 2 \left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}.$$

For Y:

$$E[Y] = \int_0^1 y f_Y(y)\, dy = \int_0^1 y(2y)\, dy = 2 \int_0^1 y^2\, dy = 2 \left[\frac{y^3}{3}\right]_0^1 = \frac{2}{3}.$$

DRAFT

Compute $E[XY]$:

Now we compute $E[XY]$ using the joint PDF:

$$\begin{aligned} E[XY] &= \int_0^1 \int_0^1 xy f_{X,Y}(x,y)\,dx\,dy \\ &= \int_0^1 \int_0^1 xy(4xy)\,dx\,dy \\ &= 4\int_0^1 \int_0^1 x^2y^2\,dx\,dy \\ &= 4\left(\int_0^1 x^2\,dx\right)\left(\int_0^1 y^2\,dy\right) \\ &= 4\cdot\frac{1}{3}\cdot\frac{1}{3} \\ &= \frac{4}{9}. \end{aligned}$$

Compute the Covariance:

Now, we compute the covariance:

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{4}{9} - \left(\frac{2}{3}\cdot\frac{2}{3}\right) = \frac{4}{9} - \frac{4}{9} = 0.$$

Thus, the covariance is:

$$\mathrm{Cov}(X,Y) = 0.$$

(c). Find the Correlation $\mathrm{Corr}(X,Y)$:

The correlation is given by:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Since $\mathrm{Cov}(X,Y) = 0$, the correlation is:

$$\mathrm{Corr}(X,Y) = 0.$$

□

5.8.11 Linear Combinations of Random Variables

When dealing with two random variables, X_1 and X_2, it is often beneficial to analyze the random variable formed by their sum. A general principle states that:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

DRAFT

This means the expected value of the sum of two random variables is equal to the sum of their individual expected values.

In addition:

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

Note that if the two random variables are **independent**, their covariance is zero, simplifying the variance of their sum to the sum of their variances:

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

Thus, the variance of the sum of two independent random variables is equal to the sum of their individual variances.

These results are straightforward, but it's crucial to remember that while the expected value of the sum of two random variables always equals the sum of their expected values, the variance of the sum only equals the sum of their variances if the random variables are independent.

Sums of Random Variables: If X_1 and X_2 are two random variables, then:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

and

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

If X_1 and X_2 are independent random variables such that $\text{Cov}(X_1, X_2) = 0$, then:

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

Now, consider a sequence of random variables $X_1, \ldots, X_n$ along with constants $a_1, \ldots, a_n$ and b. Define a new random variable Y as the linear combination:

$$Y = a_1X_1 + \cdots + a_nX_n + b$$

Linear combinations of random variables are important in various contexts, and deriving general results for them is useful. The expectation of the linear combination is:

$$E(Y) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$

which is simply the linear combination of the expectations of the random variables X_i. Additionally, if the random variables $X_1, \ldots, X_n$ are independent, then:

$$\text{Var}(Y) = a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n)$$

Note that the constant b does not affect the variance of Y, and the coefficients a_i are squared in this expression.

Draft

Theorem 5.3. *If $X_1, \ldots, X_n$ is a sequence of random variables and $a_1, \ldots, a_n$ and b are constants, then*

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b.$$

If, in addition, the random variables are independent, then

$$Var(a_1X_1 + \cdots + a_nX_n + b) = a_1^2\,Var(X_1) + \cdots + a_n^2\,Var(X_n).$$

Problem 5.27. *Suppose that $X_1, \ldots, X_n$ is a sequence of independent random variables each with an expectation μ and a variance σ^2. Consider the sample mean $\bar{X}$ defined as:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Find the mean and variance of sample mean $\bar{X}$.

Solution

Mean of the Sample Mean: Using the linearity of expectation:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

Variance of the Sample Mean: Since the X_i are independent and each has a variance σ^2:

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Therefore, the mean and variance of the sample mean $\bar{X}$ are:

$$E(\bar{X}) = \mu$$

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

DRAFT

Problem 5.28. *Let X_1 and X_2 represent the scores on two tests, with the following information:*

$$E(X_1) = 18, \quad Var(X_1) = 24, \quad E(X_2) = 30, \quad Var(X_2) = 60.$$

The scores are standardized as:

$$Y_1 = \frac{10}{3}X_1, \quad Y_2 = \frac{5}{3}X_2 + \frac{50}{3}.$$

The final score is:

$$Z = \frac{2}{3}Y_1 + \frac{1}{3}Y_2.$$

(a). Calculate the expected value of the final score $E(Z)$.

(b). Calculate $Var(Z)$ and the standard deviation σ_Z, assuming X_1 and X_2 are independent.

Solution

Let X_1 and X_2 represent the scores on two tests with the following data:

$$E(X_1) = 18, \quad \text{Var}(X_1) = 24, \quad E(X_2) = 30, \quad \text{Var}(X_2) = 60.$$

The scores are standardized as:

$$Y_1 = \frac{10}{3}X_1, \quad Y_2 = \frac{5}{3}X_2 + \frac{50}{3}.$$

The final score is:

$$Z = \frac{2}{3}Y_1 + \frac{1}{3}Y_2.$$

(a) Calculate $E(Z)$:

$$E(Z) = \frac{2}{3}E(Y_1) + \frac{1}{3}E(Y_2).$$

$$E(Y_1) = \frac{10}{3} \times 18 = 60, \quad E(Y_2) = \frac{5}{3} \times 30 + \frac{50}{3} = 66.67.$$

$$E(Z) = \frac{2}{3} \times 60 + \frac{1}{3} \times 66.67 = 62.22.$$

(b) Calculate $\text{Var}(Z)$ and σ_Z:

$$\text{Var}(Z) = \left(\frac{2}{3}\right)^2 \text{Var}(Y_1) + \left(\frac{1}{3}\right)^2 \text{Var}(Y_2).$$

$$\text{Var}(Y_1) = \left(\frac{10}{3}\right)^2 \times 24 = 266.67, \quad \text{Var}(Y_2) = \left(\frac{5}{3}\right)^2 \times 60 = 166.67.$$

DRAFT

$$\text{Var}(Z) = \frac{4}{9} \times 266.67 + \frac{1}{9} \times 166.67 = 137.04.$$

$$\sigma_Z = \sqrt{137.04} = 11.71.$$

Hence,

$$E(Z) = 62.22, \quad \text{Var}(Z) = 137.04, \quad \sigma_Z = 11.71.$$

5.8.12 Exercises

1. Suppose X, taking the values 1, 2, 3, and 4, is the service time in hours taken at Bashundhara residential area, and the Y, taking the values 1, 2, and 3, is the number of air conditioner (AC) units at the same location. The joint probability of X and Y are presented in the Table 5.17.

Table 5.17: Joint probability table

$p(x,y)$		X = number of service time (hrs)			
		1	2	3	4
Y = number of AC units	1	0.12	0.08	0.07	0.05
	2	0.08	k	0.21	0.13
	3	0.01	0.01	0.02	0.07

   (a) What is the value of k?

   (b) Find the marginal distribution of X and Y.

   (c) Find the conditional distribution of $P(X|Y = 2)$ and compute its mean.

   (d) Compute $E(XY)$.

2. Consider a random variable X denotes the **Number of Incoming Calls per Hour** (taking values 0, 1, and 2), and another random variable Y represents the **Number of Issues Resolved per Hour** (also taking values 0, 1, and 2). The joint probability mass function $p(x, y)$ for these variables is given in the table below:

DRAFT

Table 5.18: Joint Probability Mass Function $p(x, y)$ for Incoming Calls (X) and Issues Resolved (Y)

		Issues Resolved (Y)		
		0	**1**	**2**
Incoming Calls (X)	**0**	0.05	0.00	0.00
	1	0.10	0.25	0.05
	2	0.05	0.20	0.30

(a). Find the covariance of X and Y.

(b). Find the correlation of X and Y.

3. Suppose that the random variables X, Y, and Z are independent with $E(X) = 3$, $\text{Var}(X) = 4$, $E(Y) = -4$, $\text{Var}(Y) = 2$, $E(Z) = 7$, and $\text{Var}(Z) = 7$. Calculate the expectation and variance of the following random variables.

   (a) $3X + 7$
   (b) $5X - 9$
   (c) $2X + 6Y$
   (d) $4X - 3Y$
   (e) $5X - 9Z + 8$
   (f) $-3Y - Z - 5$
   (g) $X + 2Y + 3Z$
   (h) $6X + 2Y - Z + 16$

4. Suppose that items from a manufacturing process are subject to three separate evaluations, and that the results of the first evaluation X_1 have a mean value of 59 with a standard deviation of 10, the results of the second evaluation X_2 have a mean value of 67 with a standard deviation of 13, and the results of the third evaluation X_3 have a mean value of 72 with a standard deviation of 4. In addition, suppose that the results of the three evaluations can be taken to be independent of each other.

   (a) If a final evaluation score is obtained as the average of the three evaluations $X = \frac{X_1+X_2+X_3}{3}$, what are the mean and the standard deviation of the final evaluation score?

   (b) If a final evaluation score is obtained as the weighted average of the three evaluations $X = 0.4X_1 + 0.4X_2 + 0.2X_3$, what are the mean and the standard deviation of the final evaluation score?

DRAFT

5. A machine part is assembled by fastening two components of type A and one component of type B end to end. Suppose that the lengths of components of type A have an expectation of 37.0 mm and a standard deviation of 0.7 mm, whereas the lengths of components of type B have an expectation of 24.0 mm and a standard deviation of 0.3 mm. What are the expectation and variance of the length of the machine part?

6. A product is assembled by linking four components of type C and one component of type D sequentially. The lengths of components of type C have an average of 50.0 mm and a standard deviation of 0.8 mm, while the lengths of components of type D have an average of 20.0 mm and a standard deviation of 0.4 mm. Determine the average and variance of the length of the product.

7. A system is constructed by connecting five components of type G and two components of type H end to end. Assume that the lengths of components of type G have an expected value of 40.0 mm and a standard deviation of 1.5 mm, and the lengths of components of type H have an expected value of 22.0 mm and a standard deviation of 0.7 mm. What are the expectation and variance of the total length of the system?

8. A person's cholesterol level C can be measured by three different tests. Test-α returns a value X_α with a mean C and a standard deviation of 1.2, test-β returns a value X_β with a mean C and a standard deviation of 2.4, and test-γ returns a value X_γ with a mean C and a standard deviation of 3.1. Suppose that the three test results are independent. If a doctor decides to use the weighted average $0.5X_\alpha + 0.3X_\beta + 0.2X_\gamma$, what is the standard deviation of the cholesterol level obtained by the doctor?

9. Suppose that the impurity levels of water samples taken from a particular source are independent with a mean value of 3.87 and a standard deviation of 0.18.

   (a) What are the mean and the standard deviation of the sum of the impurity levels from two water samples?

   (b) What are the mean and the standard deviation of the sum of the impurity levels from three water samples?

   (c) What are the mean and the standard deviation of the average of the impurity levels from four water samples?

   (d) If the impurity levels of two water samples are averaged, and the result is subtracted from the impurity level of a third sample, what are the mean and the standard deviation of the resulting value?

DRAFT

5.9 Python Functions for Statistical Distributions

In the analysis of statistical distributions, Python provides a variety of functions to work with different types of distributions. These functions can be used to perform tasks such as generating random variates, computing probability mass functions (PMF), cumulative density functions (CDF), and more. The following Table 5.19 summarizes some of the key functions available for discrete and continuous distributions in Python.

These functions are typically part of the `'scipy.stats'` module, which includes a wide range of probability distributions and statistical functions. The table below lists each function along with a brief explanation of its purpose:

DRAFT

Python function	**Function explanation**
rvs(p, loc=0, size=1)	Random variates.
pmf(x, p, loc=0)	Probability mass function.
logpmf(x, p, loc=0)	Log of the probability mass function.
cdf(x, p, loc=0)	Cumulative density function.
logcdf(x, p, loc=0)	Log of the cumulative density function.
sf(x, p, loc=0)	Survival function (1-cdf — sometimes more accurate).
logsf(x, p, loc=0)	Log of the survival function.
ppf(q, p, loc=0)	Percent point function (inverse of cdf — percentiles).
isf(q, p, loc=0)	Inverse survival function (inverse of sf).
stats(p, loc=0, moments='mv')	Mean ('m'), variance ('v'), skew ('s'), and/or kurtosis ('k').
entropy(p, loc=0)	(Differential) entropy of the RV.
expect(func, p, loc=0, lb=None, ub=None, conditional=False)	Expected value of a function (of one argument) with respect to the distribution.
median(p, loc=0)	Median of the distribution.
mean(p, loc=0)	Mean of the distribution.
var(p, loc=0)	Variance of the distribution.
std(p, loc=0)	Standard deviation of the distribution.
interval(alpha, p, loc=0)	Endpoints of the range that contains alpha percent of the distribution.

Table 5.19: Summary of Python functions for statistical distributions.

5.10 Concluding Remarks

In this chapter, we have established a comprehensive foundation for understanding random variables and their fundamental properties. We began by defining random variables and distinguishing between discrete and continuous types, examining their respective probability functions and cumulative distribution functions. We then delved into the crucial concepts of expectation and variance, illustrating their applications through various examples. The discussion on

DRAFT

Chebyshev's Inequality highlighted its utility in providing probabilistic bounds without assuming a specific distribution. Additionally, we explored jointly distributed random variables, emphasizing the importance of understanding independence, covariance, and correlation. With these essential concepts and tools in place, we are now well-equipped to explore specific discrete probability distributions in the following chapter, where we will extend our knowledge to model and analyze discrete data more effectively.

5.11 Chapter Exercises

1. A study records the number of adverse reactions to a new drug among a group of patients. The probability distribution of the number of adverse reactions is given by:

$$P(X = x) = \begin{cases} 0.5 & \text{if } x = 0 \\ 0.3 & \text{if } x = 1 \\ 0.2 & \text{if } x = 2 \end{cases}$$

   (a) Find the expected number of adverse reactions.

   (b) Calculate the variance and standard deviation of the number of adverse reactions.

2. Let the lifetime T in hours of a certain type of electronic device have the probability density function

$$f_T(t) = \begin{cases} \frac{1}{100} e^{-\frac{t}{100}} & \text{for } t \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

   Find the expectation and variance of the lifetime.

3. Let the height H in centimeters of a particular species of plant have the probability density function

$$f_H(h) = \begin{cases} \frac{3}{64}(h - 120)^2 & \text{for } 120 \leq h \leq 124 \\ 0 & \text{elsewhere} \end{cases}$$

   Calculate the cumulative distribution function $F_H(h)$ and find the probability that a plant's height is between 121 and 123 cm.

   (a) Calculate the cumulative distribution function $F_H(h)$ and find the probability that a plant's height is between 121 and 123 cm.

   (b) Calculate the expectation and variance of H.

   (c) Find the probability that a plant's height is exactly 122 cm.

DRAFT

4. Let the length L in meters of a certain type of fish have the probability density function

$$f_L(l) = \begin{cases} 0.5l & \text{for } 0 \leq l \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the expectation, variance, and cumulative distribution function $F_L(l)$ of the length.

5. Suppose that the battery failure time, measured in hours, has a probability density function given

$$f(x) = \frac{2}{(x+1)^3}$$

for $x \geq 0$.

(a) Find the expected battery failure time.

(b) What is the probability that the battery fails within the first 4 hours.

(c) Find the median of the battery failure times.

6. Let the temperature X in degrees Fahrenheit of a particular chemical reaction with density

$$f_X(x) = \frac{x - 190}{3600} \quad 220 \leq x \leq 280.$$

(a) Find the cumulative distribution function $F(x)$ of the the temperature and it's median.

(b) Find the expectation and standard deviation of the temperature.

(c) Find the the expectation and variance of Y where $Y = \frac{5}{9}X - \frac{160}{9}$.

7. The random variable X measures the concentration of ethanol in a chemical solution, and the random variable Y measures the acidity of the solution. They have a joint probability density function

$$f(x, y) = A(20 - x - 2y)$$

for $0 \leq x \leq 5$ and $0 \leq y \leq 5$ and $f(x, y) = 0$ elsewhere.

(a) What is the value of A? What is $P(1 \leq X \leq 2, 2 \leq Y \leq 3)$?

(b) Construct the marginal probability density function $f_X(x)$.

(c) What are the expectation and the variance of the ethanol concentration?

DRAFT

Chapter 6

Some Discrete Probability Distributions

DRAFT

6.1 Introduction

In the field of data science, understanding discrete probability distributions is crucial for analyzing and modeling data that can be categorized into distinct outcomes. These distributions help data scientists interpret and predict the likelihood of various events based on historical data, which can be essential for making informed decisions and developing predictive models.

This chapter focuses on Six fundamental discrete probability distributions: the Bernoulli distribution, the Binomial distribution, the Poisson distribution, the Geometric Distribution, the Negative Binomial Distribution, and the Discrete Uniform Distribution. Each of these distributions plays a vital role in data science applications, ranging from binary classification problems to event counting and rate modeling.

Throughout this chapter, we will delve into each distribution's mathematical properties, including expected value, variance, moment generating function, and characteristic function. We will also present practical examples and exercises to illustrate how these distributions can be applied to real-world data science problems.

6.2 Bernoulli Distribution

Consider a simple experiment where we flip a fair coin. The outcome of this experiment can be either "Heads" or "Tails." We can assign a value of 1 to "Heads" and 0 to "Tails." This experiment is an example of a Bernoulli trial, which is a random experiment with exactly two possible outcomes. It is named

after **Jacob Bernoulli**, a Swiss mathematician.

Suppose we are interested in modeling the probability of getting "Heads" in a single coin flip. If the coin is fair, the probability of getting "Heads" (success) is $p = 0.5$ and the probability of getting "Tails" (failure) is $1-p = 0.5$. However, in general, the probability of success in a Bernoulli trial can be any value p such that $0 \leq p \leq 1$.

Definition: A random variable X is said to have a Bernoulli distribution with parameter p if it takes the value 1 with probability p and the value 0 with probability $1 - p$. The probability mass function (pmf) of X is given by:

$$P(X = k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

or more compactly, if $X \sim \text{Bernoulli}(0.1)$ then the pmf is

$$P(X = k) = p^k(1 - p)^{1-k} \quad \text{for } k \in \{0, 1\}.$$

6.2.1 Expected Value

The mean or expected value $E(X)$ of a Bernoulli distributed random variable X can be calculated as follows:

$$E(X) = \sum_k k \cdot P(X = k)$$

For a Bernoulli random variable X:

$$E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0)$$

Since $P(X = 1) = p$ and $P(X = 0) = 1 - p$, we have:

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

So, the mean of a Bernoulli distribution is:

$$E(X) = p$$

6.2.2 Variance

The variance $\text{Var}(X)$ of a Bernoulli distributed random variable X is defined as:

$$\text{Var}(X) = E\left[(X - E(X))^2\right] = E(X^2) - [E(X)]^2$$

First, we calculate $E(X^2)$:

DRAFT

$$E(X^2) = \sum_k k^2 \cdot P(X = k)$$

For a Bernoulli random variable X:

$$E(X^2) = 1^2 \cdot P(X = 1) + 0^2 \cdot P(X = 0) = p$$

Now, using the formula for variance:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Substitute the values we have calculated:

$$\text{Var}(X) = p - p^2 = p(1 - p)$$

So, the variance of a Bernoulli distribution is:

$$\text{Var}(X) = p(1 - p)$$

Properties

- **Mean:** The expected value (mean) of a Bernoulli random variable X is given by:
$$E(X) = p$$
- **Variance:** The variance of a Bernoulli random variable X is given by:
$$\text{Var}(X) = p(1 - p)$$
- **Standard Deviation:** The standard deviation of a Bernoulli random variable X is:
$$\sigma = \sqrt{p(1 - p)}$$

Problem 6.1. *A factory produces light bulbs, and each light bulb is tested for quality. The probability that a light bulb is defective is $p = 0.1$. Let X be a random variable that represents whether a randomly selected light bulb is defective (1 if defective, 0 if not defective).*

(a). What is the probability that a randomly selected light bulb is defective?

(b). What is the probability that a randomly selected light bulb is not defective?

(c). Compute the expected value and variance of X.

Solution

Let X follow a Bernoulli distribution with parameter $p = 0.1$, i.e., $X \sim$ Bernoulli(0.1).

(a). The probability that a randomly selected light bulb is defective is given by $P(X = 1)$:

$$P(X = 1) = p = 0.1$$

(b). The probability that a randomly selected light bulb is not defective is given by $P(X = 0)$:

$$P(X = 0) = 1 - p = 1 - 0.1 = 0.9$$

(c). For a Bernoulli random variable X with parameter p:

- The expected value $E[X]$ is:

$$E[X] = p = 0.1$$

- The variance $\text{Var}(X)$ is:

$$\text{Var}(X) = p(1-p) = 0.1 \times (1 - 0.1) = 0.1 \times 0.9 = 0.09$$

□

Problem 6.2. *We are rolling a standard six-sided die and define "success" as rolling a '3'. We need to find the probability mass function, expected value, and variance for this experiment.*

Solution

Define a random variable X where:

$$X = \begin{cases} 1 & \text{if we roll a '3' (success)} \\ 0 & \text{if we roll anything else (failure)} \end{cases}$$

This is a **Bernoulli random variable** with parameter $p = \frac{1}{6}$.
The probability mass function is:

$$P(X = x) = \begin{cases} \frac{5}{6} & \text{if } x = 0 \\ \frac{1}{6} & \text{if } x = 1 \end{cases}$$

DRAFT

Expected Value

For a Bernoulli random variable with parameter p:

$$E[X] = p = \frac{1}{6}$$

Alternatively, calculating directly:

$$\begin{aligned} E[X] &= \sum_x x \cdot P(X = x) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot \frac{5}{6} + 1 \cdot \frac{1}{6} \\ &= \frac{1}{6} \approx 0.167 \end{aligned}$$

Variance

For a Bernoulli random variable:

$$\text{Var}(X) = p(1-p) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$$

Alternatively, using the formula $\text{Var}(X) = E[X^2] - (E[X])^2$:
First, calculate $E[X^2]$:

$$\begin{aligned} E[X^2] &= \sum_x x^2 \cdot P(X = x) \\ &= 0^2 \cdot \frac{5}{6} + 1^2 \cdot \frac{1}{6} \\ &= \frac{1}{6} \end{aligned}$$

Then,

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{1}{6} - \left(\frac{1}{6}\right)^2 = \frac{5}{36} \approx 0.139$$

6.2.3 Moment Generating Function

The moment generating function $M_X(t)$ of a random variable X is defined as:

$$M_X(t) = E(e^{tX})$$

For a Bernoulli distributed random variable X with probability p of success

DRAFT

(i.e., $X = 1$ with probability p and $X = 0$ with probability $1 - p$):

$$\begin{aligned} M_X(t) = E(e^{tX}) &= e^{t\cdot 0} \cdot P(X = 0) + e^{t\cdot 1} \cdot P(X = 1) \\ &= e^0 \cdot (1 - p) + e^t \cdot p \\ &= 1 \cdot (1 - p) + e^t \cdot p \\ &= (1 - p) + pe^t \end{aligned}$$

So, the moment generating function of a Bernoulli distributed random variable X is:

$$M_X(t) = 1 - p + pe^t$$

6.2.4 Characteristic Function

The characteristic function $\varphi_X(t)$ of a Bernoulli distributed random variable X is defined as:

$$\begin{aligned} \varphi_X(t) = E(e^{itX}) &= e^{it\cdot 0} \cdot P(X = 0) + e^{it\cdot 1} \cdot P(X = 1) \\ &= e^{i\cdot 0} \cdot (1 - p) + e^{it} \cdot p \\ &= 1 \cdot (1 - p) + e^{it} \cdot p \\ &= (1 - p) + pe^{it} \end{aligned}$$

6.2.5 Probability Generating Function

For a Bernoulli random variable X with parameter p, the probability generating function (PGF) is given by:

$$G_X(s) = E[s^X]$$

Since X can take values 0 and 1, we have:

$$\begin{aligned} G_X(s) = E[s^X] &= \sum_k s^k \cdot P(X = k) \\ &= (1 - p) \cdot s^0 + p \cdot s^1 \\ &= (1 - p) + p \cdot s \end{aligned}$$

Therefore, the PGF of a Bernoulli random variable X with parameter p is:

$$G_X(s) = 1 - p + p \cdot s$$

DRAFT

6.2.6 Applications

The Bernoulli distribution is used to model binary outcomes in various scenarios, such as:

- **Quality Control in Manufacturing**

  In manufacturing, the Bernoulli distribution is used to model the probability of a defect in a production process. For example, a factory producing electronic components might use the Bernoulli distribution to determine the likelihood that a randomly selected component is defective.

- **Clinical Trials in Medicine**

  The Bernoulli distribution can model the outcome of a clinical trial for a new drug, where $X = 1$ represents a successful treatment (e.g., patient recovery) and $X = 0$ represents an unsuccessful treatment. This helps in estimating the effectiveness of the drug.

- **A/B Testing in Marketing**

  In digital marketing, A/B testing is used to compare two versions of a webpage or advertisement. The Bernoulli distribution models the probability of a user clicking on an ad or making a purchase, where $X = 1$ indicates a click or purchase and $X = 0$ indicates no click or purchase.

- **Sports Performance Analysis**

  The Bernoulli distribution can be applied to model the probability of a successful outcome in sports, such as a basketball player making a free throw or a soccer player scoring a penalty kick. Here, $X = 1$ represents a successful attempt, and $X = 0$ represents a failure.

- **Insurance Risk Assessment**

  In insurance, the Bernoulli distribution is used to model the occurrence of certain events, such as accidents or claims. For instance, $X = 1$ could represent a policyholder filing a claim within a year, and $X = 0$ could represent no claim filed.

- **Genetics**

  The Bernoulli distribution is used in genetics to model the inheritance of a particular gene. For example, $X = 1$ might represent the presence of a specific gene in an offspring, and $X = 0$ represents its absence, assuming a certain probability of inheritance.

DRAFT

6.2.7 Python Code for Bernoulli Distribution

In Python, we can calculate various characteristics of the Bernoulli distribution using the ‘scipy.stats‘ module. Below, we detail how to compute the Probability Mass Function, Cumulative Distribution Function, Mean, Variance, and Probability Generating Function, etc.

Python Code

Here's how we can compute these characteristics using Python:

```
import numpy as np
from scipy.stats import bernoulli

# Define the parameter p
p = 0.25

# Bernoulli distribution
dist = bernoulli(p)

# 1. Probability Mass Function (PMF)
x_values = [0, 1]  # Possible values for a Bernoulli
  random variable
pmf_values = dist.pmf(x_values)
print("PMF values for x = 0 and x = 1:", pmf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist.cdf(x_values)
print("CDF values for x = 0 and x = 1:", cdf_values)

# 3. Mean (Expected Value)
mean = dist.mean()
print("Mean (Expected Value):", mean)

# 4. Variance
variance = dist.var()
print("Variance:", variance)

# 5. Probability Generating Function (PGF)
def pgf(t, p):
  return (1 - p) + p * t

# PGF values at t = 0 and t = 1
pgf_values = [pgf(t, p) for t in [0, 1]]
print("PGF values for t = 0 and t = 1:", pgf_values)
```

Explanations

- **Probability Mass Function (pmf):**

- The pmf gives the probability of each outcome (0 or 1). Use `dist.pmf(x_values)` to compute these probabilities.

- **Cumulative Distribution Function (CDF):**
  - The CDF gives the cumulative probability up to each outcome. Use `dist.cdf(x_values)` to compute these values.

- **Mean (Expected Value):**
  - The mean is simply the parameter p of the Bernoulli distribution. Use `dist.mean()` to get this value.

- **Variance:**
  - The variance of a Bernoulli distribution is $p(1-p)$. Use `dist.var()` to compute this.

- **Probability Generating Function (PGF):**
  - The PGF is calculated using the formula $G_X(t) = 1 - p + pt$. Define a function `pgf(t, p)` to compute PGF values for specific t values (e.g., 0 and 1).

6.2.8 Exercises

1. Let $X \sim \text{Bernoulli}(0.4)$. Calculate the following quantities:
   (a) $P(X = 1)$
   (b) $E[X]$
   (c) $\text{Var}(X)$

2. A medical test for a certain disease has a 98% chance of correctly identifying a diseased person (true positive) and a 2% chance of incorrectly identifying a healthy person as diseased (false positive). Define a random variable X such that $X = 1$ if the test result is positive (either true positive or false positive) and $X = 0$ if the test result is negative.
   (a) What is the probability that a randomly selected test result is positive?
   (b) What is the expected value (mean) of X?
   (c) What is the variance of X?

3. A coin is biased such that the probability of heads is 0.6. Let X be a random variable where $X = 1$ if heads appears and $X = 0$ if tails appears.
   (a) What is the distribution of X? State the parameter.

DRAFT

(b) Find $E[X]$ and $\text{Var}(X)$.

(c) Find $E[3X + 2]$ and $\text{Var}(3X + 2)$.

4. Given that $X \sim \text{Bernoulli}(p)$ and $\text{Var}(X) = 0.21$:

   (a) Find all possible values of p.

   (b) For each value of p, calculate $E[X]$.

5. A basketball player has a free throw success rate of 70%. Define 'success' as making the free throw. What is the probability mass function, expected value, and variance for a single free throw attempt?

6. A quality control inspector examines a product, which has a 5% defect rate. Define 'success' as finding a defective product. What is the probability mass function, expected value, and variance for inspecting a single product?

7. A diagnostic test has a 92% chance of correctly detecting a condition when it is present (true positive rate) and an 8% chance of detecting the condition when it is not present (false positive rate). Define a random variable X such that $X = 1$ if the test result is positive (either true positive or false positive) and $X = 0$ if the test result is negative.

   (a) What is the probability that a randomly selected test result is positive?

   (b) What is the expected value (mean) of X?

   (c) What is the variance of X?

   (d) In a group of 100 people who took the test, what is the expected number of positive test results?

DRAFT

6.3 Binomial Distribution

Consider the simple experiment of flipping a fair coin, where the outcome can be either "Heads" (success) or "Tails" (failure). We assign a value of 1 to "Heads" and 0 to "Tails," making this a Bernoulli trial. Suppose we flip the coin n times. We are interested in modeling the probability of obtaining exactly k "Heads" (successes) in n flips.

This problem is an example of a Binomial experiment, which consists of n independent Bernoulli trials, each with the same probability of success p. To calculate the probability of getting exactly k successes in n trials, consider a specific sequence of trials where exactly k trials are successes and $n - k$ trials are failures. The probability of such a specific sequence is given by

$$p^k \times (1 - p)^{n-k}$$

where p is the probability of success and $1 - p$ is the probability of failure.

We need to account for all possible sequences of n trials that result in exactly k successes. The number of ways to choose k positions for successes out of n positions is given by the Binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $\binom{n}{k}$ represents the number of combinations of n items taken k at a time.

The total probability of having exactly k successes in n trials is the product of the probability of any specific sequence and the number of such sequences. Thus, the probability mass function of the Binomial distribution is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, n,$$

where X is the random variable representing the number of successes.

The Binomial distribution is a discrete probability distribution. This distribution has the following conditions:

1. **Fixed Number of Trials:** The experiment is repeated a fixed number of times, denoted as n.
2. **Two Possible Outcomes:** Each trial results in one of two outcomes, often referred to as "success" and "failure."
3. **Constant Probability:** The probability of success in each trial is constant and denoted by p. Consequently, the probability of failure is $1 - p$.
4. **Independence:** The outcome of one trial is independent of the outcomes of other trials.

If these conditions are met, the number of successes in n trials follows a Binomial distribution with parameters n and p.

Definition: A random variable X that represents the number of successes in n independent Bernoulli trials, each with a probability of success p, is said to follow a Binomial distribution if its probability mass function (pmf) is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, n,$$

where $\binom{n}{k}$ is the Binomial coefficient. We write this as

$$X \sim \mathrm{B}(n, p).$$

DRAFT

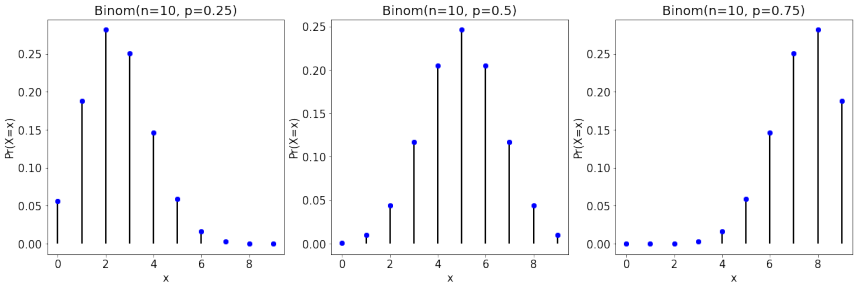


Figure 6.1: Graphical Presentation of pmf of Binomial Distribution.

The graphical presentation of the Binomial distribution for $n = 10$ and $p = 0.25$, $p = 0.5$, and $p = 0.75$ is depicted in Figure 6.1. The Python code used to generate these figures is provided below.

```python
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import binom

# Parameters
n = 10
p = 0.5

# Values
k = np.arange(0, n+1)
pmf = binom.pmf(k, n, p)

# Plot
plt.bar(k, pmf, color='blue', edgecolor='black')
plt.xlabel('Number of Successes')
plt.ylabel('Probability')
plt.title('Binomial PMF (n=10, p=0.5)')
plt.xticks(k)
plt.grid(True)
plt.show()
```

Symmetric Binomial Distributions A $\mathrm{B}(n, 0.5)$ distribution is a symmetric probability distribution for any value of the parameter n. The distribution is symmetric about the expected value $n/2$.

6.3.1 Expected Value

Let $X \sim \mathrm{B}(n, p)$. The expected value $E[X]$ is

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \cdot P(X = k) \\ &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad \left[\text{Since, } k \cdot \binom{n}{k} = n \cdot \binom{n-1}{k-1}\right] \\ &= n \cdot p \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \end{aligned}$$

Using the binomial expansion of $(p + (1-p))^{n-1}$, we have

$$\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = 1$$

Thus,

$$E(X) = n \cdot p \cdot 1 = np$$

Therefore, the expected value of X is

$$E(X) = np$$

Alternatively:

By recognizing that X can be written as the sum of n independent Bernoulli trials X_i

$$X = X_1 + X_2 + \cdots + X_n$$

where each X_i is an indicator variable such that

$$X_i = \begin{cases} 1, & \text{if trial } i \text{ is a success,} \\ 0, & \text{otherwise.} \end{cases}$$

Each $X_i \sim \text{Bernoulli}(p)$ with

$$E[X_i] = p$$

Using the linearity of expectation, we have

$$E[X] = E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n] = n \cdot p$$

Thus, the expected value $E[X]$ is

$$E[X] = np$$

DRAFT

6.3.2 Variance and Standard Deviation

The variance is defined as

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

We use the identity:

$$E[X(X-1)] = \sum_{k=0}^{n} k(k-1) \cdot \binom{n}{k} p^k (1-p)^{n-k}$$

Since $X \sim \mathrm{B}(n,p)$ be a binomial random variable, we can write

$$X = \sum_{i=1}^{n} X_i,$$

where each X_i is an indicator variable such that

$$X_i = \begin{cases} 1, & \text{if trial } i \text{ is a success,} \\ 0, & \text{otherwise.} \end{cases}$$

Since the X_i are independent Bernoulli random variables with parameter p, we have

$$E[X_i] = p \quad \text{and} \quad E[X_i X_j] = E[X_i]E[X_j] = p^2 \quad \text{for } i \neq j.$$

Then,

$$\begin{aligned} E[X(X-1)] &= E\left[\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} X_i X_j\right] = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} E[X_i X_j] \\ &= \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} p^2 = n(n-1)p^2. \end{aligned}$$

Therefore, we have

$$E[X(X-1)] = n(n-1)p^2.$$

Then,

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= E[X(X-1)] + E[X] - (E[X])^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p) \end{aligned}$$

DRAFT

Therefore,

$$\text{Var}(X) = np(1-p)$$

Alternatively:

Since $X_1, X_2, \ldots, X_n$ are independent random variables, we can use Theorem 5.3. Thus, the variance $\text{Var}(X)$ can be computed as follows

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= \text{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) \\ &= np(1-p).\end{aligned}$$

The standard deviation σ is

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{np(1-p)}$$

Properties

- **Mean:** The expected value (mean) of a Binomial random variable X is given by
$$E(X) = np$$
- **Variance:** The variance of a Binomial random variable X is given by
$$\text{Var}(X) = np(1-p)$$
- **Standard Deviation:** The standard deviation of a Binomial random variable X is
$$\sigma = \sqrt{np(1-p)}$$

6.3.3 Example

Let us consider a biased coin where the probability of getting "Heads" is $p = 0.7$. Suppose we flip this coin $n = 10$ times. The random variable X representing the number of "Heads" in 10 flips follows a Binomial distribution with parameters $n = 10$ and $p = 0.7$.

The pmf of X is

$$P(X = k) = \binom{10}{k}(0.7)^k(0.3)^{10-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, 10.$$

The mean and variance of X are

$$E(X) = 10 \times 0.7 = 7$$

$$\text{Var}(X) = 10 \times 0.7 \times (1 - 0.7) = 2.1$$

Thus, we can model and analyze the number of "Heads" in 10 coin flips using the Binomial distribution.

Problem 6.3. *A factory produces light bulbs, and 5% of them are defective. Suppose a quality control inspector randomly selects 8 bulbs for testing.*

(a). What is the probability that exactly 2 of the 8 bulbs are defective?

(b). What is the probability that at most 2 bulbs are defective?

(c). What is the probability that more than 1 bulb is defective?

(d). What is the probability that at least 1 bulb is defective?

(e). What is the expected number of defective bulbs in the sample of 8?

(f). What is the variance and standard deviation of the number of defective bulbs in the sample?

Solution

Let X be the number of defective bulbs in the sample. Here, X follows a binomial distribution with parameters $n = 8$ and $p = 0.05$. The pmf of X is

$$P(X = k) = \binom{8}{k}(0.05)^k(1 - 0.05)^{8-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, 8.$$

(a). Probability of Exactly 2 Defective Bulbs: The probability of exactly 2 defective bulbs is

$$P(X = 2) = \binom{8}{2}(0.05)^2(1 - 0.05)^{8-2} = 28 \times (0.05)^2 \times (0.95)^6 \approx 0.0514$$

Thus, the probability that exactly 2 bulbs are defective is approximately 0.0514.

(b). Probability of At Most 2 Defective Bulbs:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

Compute each term, we have

$$P(X = 0) = \binom{8}{0}(0.05)^0(0.95)^8 \approx 0.6634$$

$$P(X = 1) = \binom{8}{1}(0.05)^1(0.95)^7 \approx 0.2770$$

$$P(X = 2) = \binom{8}{2}(0.05)^2(0.95)^6 \approx 0.0514$$

Hence,

$$P(X \leq 2) \approx 0.6634 + 0.2770 + 0.0514 = 0.9918$$

Thus, the probability that at most 2 bulbs are defective is approximately 0.9918. □

DRAFT

(c). We want to find $P(X > 1)$. We have,

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - \{P(X = 0) + P(X = 1)\} \\ &\approx 1 - \{0.6634204 + 0.2793349\} \\ &= 1 - 0.9427553 \\ &= 0.0572 \end{aligned}$$

(d). We want to find $P(X \geq 1)$. It is easier to calculate this as the complement of $P(X = 0)$.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &\approx 1 - 0.6634204 \\ &= 0.3366 \end{aligned}$$

(e). For a binomial distribution, the expected value (mean) is given by

$$E(X) = np = 8 \times 0.05 = 0.4$$

The expected number of defective bulbs is 0.4.

(f). For a binomial distribution, the variance is given by $\text{Var}(X) = np(1-p)$.

$$\text{Var}(X) = 8 \times 0.05 \times (1 - 0.05) = 8 \times 0.05 \times 0.95 = 0.38$$

The standard deviation is the square root of the variance: $sd(X) = \sqrt{\text{Var}(X)}$.

$$sd(X) = \sqrt{0.38} \approx 0.61644$$

Rounding to four decimal places, the standard deviation is approximately 0.6164.

Problem 6.4. *A pharmaceutical company is testing a new drug to see if it improves recovery rates. The probability that a patient responds positively to the drug is $p = 0.25$. Suppose the company tests the drug on 10 patients. Let X be the number of patients who respond positively to the drug.*

(a). What is the probability that exactly 3 patients respond positively?

(b). What is the probability that at most 3 patients respond positively?

(c). Calculate the expected number of patients who respond positively and the variance of X.

(d). If Y represents the number of patients who do not respond positively, find the probability distribution of Y and its mean and variance.

DRAFT

Solution

Let X be the number of patients who respond positively out of 10 patients, with each patient responding positively with probability $p = 0.25$. Then X follows a binomial distribution with pmf of X is

$$P(X = k) = \binom{10}{k}(0.25)^k(1 - 0.25)^{10-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, 10.$$

(a). The probability of exactly 3 patients responding positively is given by the binomial probability mass function

$$\begin{aligned} P(X = 3) &= \binom{10}{3}(0.25)^3(0.75)^7 \\ &= 120 \times (0.25)^3 \times (0.75)^7 \approx 120 \times 0.015625 \times 0.1335 \\ &\approx 0.2503 \end{aligned}$$

Therefore, the probability that exactly 3 patients respond positively is approximately 0.2503.

(b). The probability that at most 3 patients respond positively is the cumulative probability $P(X \leq 3)$, which is the sum of the probabilities for $X = 0, 1, 2, 3$. Therefore,

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \binom{10}{0}(0.25)^0(0.75)^{10} + \binom{10}{1}(0.25)^1(0.75)^9 + \binom{10}{2}(0.25)^2(0.75)^8 \\ &\quad + 0.2503 \quad \text{(from part (a))} \\ &\approx 0.0563 + 0.1877 + 0.2816 + 0.2503 = 0.7760 \end{aligned}$$

Therefore, the probability that at most 3 patients respond positively is approximately 0.7760.

(c). For a binomial distribution $X \sim \mathrm{B}(n, p)$, the expected value and variance are given by

$$E(X) = np \quad \text{and} \quad \mathrm{Var}(X) = np(1 - p)$$

Substituting $n = 10$ and $p = 0.25$,

$$E(X) = 10 \times 0.25 = 2.5$$

$$\mathrm{Var}(X) = 10 \times 0.25 \times 0.75 = 1.875$$

Therefore, the expected number of patients who respond positively is 2.5 and the variance is 1.875.

DRAFT

(d). Since Y represents the number of patients who do not respond positively, we can express $Y = 10 - X$. The number of patients who do not respond positively follows a binomial distribution $Y \sim \mathrm{B}(n = 10, p = 0.75)$.

The probability mass function of Y is

$$P(Y = k) = \binom{10}{k}(0.75)^k(0.25)^{10-k}, \quad k = 0, 1, 2, \ldots, 10$$

The expected value and variance of Y are

$$E(Y) = np = 10 \times 0.75 = 7.5$$

$$\mathrm{Var}(Y) = np(1-p) = 10 \times 0.75 \times 0.25 = 1.875$$

Therefore, the probability distribution of Y is binomial with parameters $n = 10$ and $p = 0.75$, and the mean and variance are $E(Y) = 7.5$ and $\mathrm{Var}(Y) = 1.875$. □

6.3.4 Python Code for Binomial Distribution

In Python, you can compute various characteristics of the Binomial distribution using the 'scipy.stats' module. Below is a demonstration of how to compute these characteristics.

Python Code

Here's how you can calculate various characteristics of a Binomial distribution

```python
import numpy as np
from scipy.stats import binom

# Define parameters
n = 10  # number of trials
p = 0.25  # probability of success

# Binomial distribution
dist = binom(n, p)

# 1. Probability Mass Function (pmf)
x_values = np.arange(0, n + 1)  # possible values for the random variable
pmf_values = dist.pmf(x_values)
print("PMF values for x = 0 to 10:", pmf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist.cdf(x_values)
print("CDF values for x = 0 to 10:", cdf_values)

# 3. Mean (Expected Value)
```

DRAFT

```
mean = dist.mean()
print("Mean (Expected Value):", mean)

# 4. Variance
variance = dist.var()
print("Variance:", variance)

# 5. Probability Generating Function (PGF)
def pgf(t, n, p):
  return (1 - p + p * t) ** n

# PGF values at t = 0 and t = 1
pgf_values = [pgf(t, n, p) for t in [0, 1]]
print("PGF values for t = 0 and t = 1:", pgf_values)
```

Explanations

- **Probability Mass Function (pmf):**
  - The pmf provides the probability of each number of successes. Compute these probabilities using `dist.pmf(x_values)`.
- **Cumulative Distribution Function (CDF):**
  - The CDF provides the cumulative probability up to a certain number of successes. Compute these values using `dist.cdf(x_values)`.
- **Mean (Expected Value):**
  - The mean is given by np. Compute this using `dist.mean()`.
- **Variance:**
  - The variance is given by $np(1-p)$. Compute this using `dist.var()`.
- **Probability Generating Function (PGF):**
  - The PGF is given by $G_X(t) = (1 - p + pt)^n$. Define a function `pgf(t, n, p)` to compute PGF values for specific t values.

6.3.5 Exercises

1. Define the Binomial distribution. Include the conditions that must be met for a random variable to follow a Binomial distribution.
2. Prove that the expected value (mean) of a Binomially distributed random variable X with parameters n and p is $E(X) = np$. Also, prove that the variance of X is given by $\text{Var}(X) = np(1 - p)$.

DRAFT

3. Prove that the sum of the probabilities of all possible outcomes of a Binomial random variable X with parameters n and p is equal to 1. That is, show that
$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

4. A fair coin is flipped 5 times. What is the probability of getting exactly 3 heads?

5. In a class, the probability that a student passes an exam is 0.8 If 15 students are randomly selected, calculate the expected number of students who pass the exam.

6. Show that
$$\binom{n}{k} = \binom{n}{n-k}$$
and use this property to demonstrate that the sum of the Binomial coefficients for a given n is symmetric around $k = \frac{n}{2}$.

7. Find the moment generating function of the Binomial distribution and use it to compute the first two moments (mean and variance).

8. A factory produces widgets, with a 1% defect rate. If a sample of 50 widgets is tested, what is the probability that at least 3 widgets are defective?

9. A medical test has a 95% sensitivity and a 90% specificity. If 10 patients are tested who all have the disease, calculate the probability that exactly 9 patients test positive.

10. A soccer player has a 60% chance of scoring a goal in a penalty kick. If the player takes 12 penalty kicks, find the probability that the player scores more than 8 goals.

11. In a marketing campaign, the probability of converting a lead into a customer is 10%. If 25 leads are contacted, find the probability of converting at most 5 leads.

12. In a clinical trial, the number of successes (patients who show improvement) follows a Binomial distribution with $n = 10$ and $p = 0.7$.
    (a) What is the probability of exactly 7 successes?
    (b) What is the probability of at least 8 successes?
    (c) Find the expected number of successes.

13. Prove the following Binomial coefficient identity
$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$
Use a combinatorial argument or algebraic manipulation to justify this identity.

DRAFT

6.4 Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, provided that these events occur with a known constant mean rate and independently of the time since the last event. It is often used to model rare events.

Unlike the Bernoulli distribution (which models a single trial) or the Binomial distribution (which models a fixed number of trials), the Poisson distribution models the number of occurrences of an event, X, over a continuous, fixed interval (such as an hour, a square meter, or a kilometer of road).

The Poisson distribution answers the question: "How many times will an event occur in a given interval?" rather than "Will an event occur or not?" The Poisson distribution is fully characterized by its single parameter, λ. For a random variable X that follows a Poisson distribution with parameter λ (denoted $X \sim \mathrm{P}(\lambda)$), the pmf is given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where

- k is the number of events,
- λ is the average rate,
- e is the base of the natural logarithm (approximately 2.71828).

A unique property of the Poisson distribution is that its mean and variance are equal. That is,

$$E[X] = \mathrm{Var}(X) = \lambda.$$

6.4.1 Conditions for the Poisson Distribution

The Poisson distribution is appropriate when the following conditions are met:

(i). **Events occur independently:** The occurrence of one event does not affect the probability of another event occurring.

(ii). **Events occur at a constant average rate:** The parameter λ represents this average rate and remains constant over the interval.

(iii). **Events are relatively rare within the interval:** More specifically, the probability of more than one event occurring in a very short interval is negligible.

DRAFT

(iv). **The probability of an event in a small interval is proportional to the length of the interval:** For very small intervals, the probability is approximately $\lambda \cdot \Delta t$, where Δt is the interval length.

When these conditions are satisfied, the number of events occurring in a fixed interval follows a Poisson distribution. It is observed that the series expansion of e^{λ} guarantees that the total probability sums to 1, as shown below:

$$\begin{aligned}\sum_{k=0}^{\infty} P(X=k) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\left(\frac{1}{0!}+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\cdots\right) \\ &= e^{-\lambda}\cdot e^{\lambda} = 1.\end{aligned}$$

Definition: A random variable X is said to follow a Poisson distribution with parameter λ, written as $X \sim \mathrm{P}(\lambda)$, if its probability mass function is given by

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, 3, \ldots.$$

The Poisson distribution is particularly useful for modeling the number of occurrences of a certain event within a specified unit of time, distance, or volume, with both its mean and variance equal to λ.

Problem 6.5. *A hospital emergency room receives an average of **4 patients per hour** during weekend nights. Let the number of patients arriving in a given time interval follow a Poisson distribution.*

(a) What is the probability that exactly 6 patients arrive in a given hour?

(b) What is the probability that fewer than 3 patients arrive in a given hour?

(c) What is the probability that at least 1 patient arrives during a 30-minute period?

Solution

Let X be the number of patients arriving at the emergency room **per hour**. Since patients arrive independently, at a constant average rate, and two arrivals cannot occur simultaneously, X follows a **Poisson distribution**:

$$X \sim \text{Poisson}(\lambda = 4)$$

The probability mass function (PMF) is:

$$P(X=k) \;=\; \frac{e^{-\lambda}\,\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

DRAFT

(a): We want the probability of *exactly* 6 arrivals in one hour with $\lambda = 4$:

$$P(X = 6) \;=\; \frac{e^{-4} \cdot 4^6}{6!} \;=\; \frac{(0.018316)(4096)}{720} = 0.1042$$

There is approximately a **10.42% chance** that exactly 6 patients will arrive at the emergency room in any given hour.

(b): Fewer than 3 patients means $X \in \{0, 1, 2\}$:

$$\begin{aligned} P(X < 3) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!} \\ &= e^{-4}\left(\frac{1}{1} + \frac{4}{1} + \frac{16}{2}\right) \\ &= e^{-4}(1 + 4 + 8) = (0.018316)(13) = 0.2381 \end{aligned}$$

There is approximately a **23.81% chance** that fewer than 3 patients arrive in a given hour.

(c): Let Y be the number of patients arriving in a **30-minute** window. Since the observation period is halved, the rate is scaled accordingly:

$$Y \;\sim\; \text{Poisson}\big(\lambda' = 4 \times \tfrac{1}{2} = 2\big)$$

Using the complement rule:

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - \frac{e^{-2} \cdot 2^0}{0!} = 1 - 0.1353 = 0.8647$$

There is approximately an **86.47% chance** that at least one patient arrives during any given 30-minute window. □

Problem 6.6. *A quality inspector at a glass manufacturing company checks each glass sheet for imperfections. Suppose the number of flaws in each sheet follows a Poisson distribution with the expected number of flaws per sheet is 0.5.*

(a). Determine the probability that a glass sheet has no flaws.

(b). Sheets with two or more flaws are scrapped by the company. Estimate the percentage of glass sheets that need to be scrapped and recycled.

Solution

Let X denote the number of flaws in a glass sheet. Given that the number of flaws follows a Poisson distribution with mean $\lambda = 0.5$, we write

$$X \sim \text{Poisson}(0.5).$$

DRAFT

(a). Probability of No Flaws: The probability that a glass sheet has no flaws is given by

$$P(X = 0) = \frac{e^{-0.5} \cdot 0.5^0}{0!} = e^{-0.5} \approx 0.607$$

Thus, approximately 61% of the glass sheets are in "perfect" condition.

(b). Probability of Two or More Flaws: The probability of having two or more flaws ($X \geq 2$) can be computed as

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-0.5} - (e^{-0.5} \cdot 0.5) \\ &\approx 1 - 0.607 - 0.305 = 0.090 \end{aligned}$$

Hence, about 9% of the glass sheets have two or more flaws and need to be scrapped.

□

The Poisson distribution is used to model the **number of rare events** occurring within a fixed interval of time or space, under the assumption that these events occur independently and at a constant average rate. This distribution helps answer questions about the likelihood of observing a specific number of events given an average rate, such as predicting the number of genetic mutations in bacterial cultures or call arrivals in a call center.

Problem 6.7. *A researcher is studying a rare genetic mutation. In previous large-scale studies, the mutation was detected in an average of 2 individuals per 100-person sample. A new random sample of 100 individuals is now selected.*

(a). Calculate the probability that exactly one individual in the sample has the mutation.

(b). Determine the probability that at least three individuals in the sample have the mutation.

(c). Estimate the percentage of samples in which two or more individuals are expected to have the mutation.

Solution

Let X denote the number of individuals (out of 100) who carry the rare mutation. It is given,

$$X \sim \text{Poisson}(\lambda = 2).$$

(a). The probability that exactly one individual has the mutation is given by

$$P(X = 1) = \frac{e^{-2} \cdot 2^1}{1!} = 2e^{-2} \approx 0.2707$$

There is about a 27% chance that exactly one person in the sample of 100 individuals will have the mutation.

DRAFT

(b). The probability of at least three individuals having the mutation is

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[e^{-2} + 2e^{-2} + \frac{2^2 \cdot e^{-2}}{2!}\right] \\ &= 1 - (0.1353 + 0.2707 + 0.2707) = 1 - 0.6767 = 0.3233 \end{aligned}$$

Approximately 32% of samples are expected to contain three or more mutation cases, even though the mutation is rare.

(c). Percentage of Samples with Two or More Individuals Having the Mutation

To estimate the percentage of samples where two or more individuals have the mutation:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)]$$

Thus,

$$P(X \geq 2) = 1 - (0.1353 + 0.2707) = 1 - 0.4060 = 0.5940$$

The percentage is

$$0.5940 \times 100\% = 59.40\%$$

That is, about 59% of all samples are expected to contain two or more individuals with the mutation.

□

6.4.2 Expected Value

The expected value $E(X)$ of a discrete random variable X is defined as

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \cdot P(X = x) \\ &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= 0 \cdot \frac{\lambda^0 e^{-\lambda}}{0!} + \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!}. \end{aligned}$$

To adjust the index, let $x' = x - 1$. Then, when x starts from 1, x' starts from 0.

DRAFT

Rewriting the sum in terms of x',

$$\begin{aligned} E(X) &= e^{-\lambda} \sum_{x'=0}^{\infty} \frac{\lambda^{x'+1}}{x'!} \\ &= \lambda e^{-\lambda} \sum_{x'=0}^{\infty} \frac{\lambda^{x'}}{x'!}. \end{aligned}$$

Recognize that the sum is the Taylor series expansion of e^{λ},

$$\sum_{x'=0}^{\infty} \frac{\lambda^{x'}}{x'!} = e^{\lambda}$$

Thus,

$$E(X) = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

6.4.3 Variance

To find the variance, we first need to calculate $E(X^2)$. We use,

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{\infty} x^2 \cdot P(X = x) \\ &= \sum_{x=0}^{\infty} x^2 \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x}{(x-1)!} \\ &= \sum_{x=1}^{\infty} \frac{x\lambda^x e^{-\lambda}}{(x-1)!} \end{aligned}$$

Change the index of summation. Let $x' = x - 1$. Therefore, $x = x' + 1$, and when x starts from 1, x' starts from 0.

Therefore, we have

$$\begin{aligned} E(X^2) &= e^{-\lambda} \sum_{x'=0}^{\infty} \frac{(x'+1)\lambda^{x'+1}}{x'!} \\ &= e^{-\lambda} \left(\sum_{x'=0}^{\infty} \frac{x'\lambda^{x'+1}}{x'!} + \sum_{x'=0}^{\infty} \frac{\lambda^{x'+1}}{x'!} \right) \\ &= e^{-\lambda} \left(\lambda \sum_{x'=0}^{\infty} \frac{x'\lambda^{x'}}{x'!} + \lambda \sum_{x'=0}^{\infty} \frac{\lambda^{x'}}{x'!} \right) \\ &= e^{-\lambda} \left(\lambda \cdot \lambda e^{\lambda} + \lambda e^{\lambda} \right) \\ &= \lambda^2 + \lambda. \end{aligned}$$

DRAFT

Now, using the definition of variance,

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= (\lambda^2 + \lambda) - \lambda^2 \\ &= \lambda. \end{aligned}$$

6.4.4 Moment Generating Function

The moment generating function (MGF) $M_X(t)$ of X is defined as,

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \sum_{x=0}^{\infty} e^{tx} \cdot P(X = x) \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} \\ &= e^{-\lambda} \cdot e^{e^t\lambda} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

Thus, the moment generating function $M_X(t)$ of a Poisson random variable X with parameter λ is,

$$M_X(t) = e^{\lambda(e^t-1)}.$$

6.4.5 Characteristic Function

The characteristic function $\phi_X(t)$ of X is defined as,

$$\begin{aligned} \phi_X(t) = E[e^{tX}] &= \sum_{x=0}^{\infty} e^{itx} \cdot P(X = x) \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^{it}\lambda)^x}{x!} \\ &= e^{-\lambda} \cdot e^{e^{it}\lambda} \\ &= e^{\lambda(e^{it}-1)}. \end{aligned}$$

Thus, the characteristic function $\phi_X(t)$ of a Poisson random variable X with parameter λ is,

$$\phi_X(t) = e^{\lambda(e^{it}-1)}.$$

6.4.6 Approximation of Binomial Distribution Using Poisson Distribution

In statistical theory, the Poisson distribution can be used as an approximation to the Binomial distribution under certain conditions. Specifically, when dealing

DRAFT

with scenarios where the number of trials n is very large (e.g., $n \geq 100$) and the probability of success p is very small (e.g., $p \leq 0.05$), while the product np remains constant (usually, $np < 5$ is used for high accuracy), the Binomial distribution can be approximated by the Poisson distribution.

Theorem 6.1. ***Approximation of Binomial Distribution by Poisson Distribution:*** *Let X be a Binomial random variable with probability distribution $B(x; n, p)$. When $n \to \infty$, $p \to 0$, and $\lambda = np$ remains constant, the Binomial distribution $B(n, p)$ converges to the Poisson distribution $P(\lambda)$ as $n \to \infty$. That is,*

$$B(n, p) \xrightarrow{n\to\infty} P(\lambda)$$

where,

$$\lambda = np$$

Problem 6.8. *A box contains 500 electrical switches, each with a probability of 0.005 of being defective. What is the probability that the box contains no more than 3 defective switches?*

*(i) Use the **Binomial Distribution**.*

*(ii) Use the **Poisson Distribution** to make an approximation.*

(iii) Compare the results.

Solution:

Given, total number of switches $n = 500$, and probability of a switch being defective $p = 0.005$. Let a random variable X represents the number of defective switches. We are required to find $P(X \leq 3)$, i.e., the probability that the number of defective switches is no more than 3.

(i). Using Binomial Distribution: The probability mass function for the **Binomial Distribution** is given by

$$P(X = k) = \binom{500}{k} 0.005^k (1 - 0.005)^{500-k} \quad \text{for} \quad k = 0, 1, 2, \ldots, 500.$$

We need to calculate $P(X \leq 3)$. Now we have,

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

where,

$$P(X = 0) = \binom{500}{0} (0.005)^0 (0.995)^{500} \approx 0.08157$$

$$P(X = 1) = \binom{500}{1} (0.005)^1 (0.995)^{499} \approx 0.20495$$

DRAFT

$$P(X = 2) = \binom{500}{2}(0.005)^2(0.995)^{498} \approx 0.25697$$

$$P(X = 3) = \binom{500}{3}(0.005)^3(0.995)^{497} \approx 0.21435$$

Hence,

$$P(X \leq 3) \approx 0.08157 + 0.20495 + 0.25697 + 0.21435 = 0.75784$$

Thus, the probability that the box contains no more than 3 defective switches using the binomial distribution is approximately 0.757 or 75.78%.

(ii). Using Poisson Distribution (Approximation) For $n = 500$ and $p = 0.005$, the parameter λ for the Poisson distribution is

$$\lambda = np = 500 \times 0.005 = 2.5$$

The probability mass function of the Poisson distribution is

$$P(X = k) = \frac{e^{-2.5}2.5^k}{k!} \quad \text{for} \quad k = 0, 1, 2, \ldots.$$

The probability that the number of defective switches X is at most 3 is

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= e^{-2.5} + 2.5e^{-2.5} + \frac{2.5^2e^{-2.5}}{2} + \frac{2.5^3e^{-2.5}}{6} \\ &= 0.08208 + 0.2052 + 0.2565 + 0.2138 \\ &\approx 0.7576 \end{aligned}$$

The probability that the box contains no more than 3 defective switches is approximately 0.7576 or 75.76%. □

(iii). Comparison: From Part (a) and Part (b), we have,

- **Binomial result**: 75.78%
- **Poisson result**: 75.76%

The Poisson approximation is extremely close to the binomial result, differing by only 0.02%. This demonstrates its effectiveness for modeling rare events when n is large and p is small. However, the Poisson distribution provides a simpler calculation when n is large and p is small.

DRAFT

6.4.7 Python Code for Poisson Distribution

The Poisson distribution models the number of occurrences of an event in a fixed interval of time or space, given a constant mean rate of occurrence. In Python, you can calculate various characteristics of the Poisson distribution using the 'scipy.stats' module. Below is a demonstration of how to compute these characteristics.

Python Code

Here's how you can calculate various characteristics of the Poisson distribution:

```
import numpy as np
from scipy.stats import poisson

# Define the parameter lambda (mean rate of occurrence)
lambda_ = 4

# Poisson distribution
dist = poisson(mu=lambda_)

# 1. Probability Mass Function (PMF)
x_values = np.arange(0, 11)  # Possible values for the
  random variable
pmf_values = dist.pmf(x_values)
print("PMF values for x = 0 to 10:", pmf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist.cdf(x_values)
print("CDF values for x = 0 to 10:", cdf_values)

# 3. Mean (Expected Value)
mean = dist.mean()
print("Mean (Expected Value):", mean)

# 4. Variance
variance = dist.var()
print("Variance:", variance)

# 5. Probability Generating Function (PGF)
def pgf(t, lambda_):
  return np.exp(lambda_ * (np.exp(t) - 1))

# PGF values at t = 0 and t = 1
pgf_values = [pgf(t, lambda_) for t in [0, 1]]
print("PGF values for t = 0 and t = 1:", pgf_values)
```

Explanations

- **Probability Mass Function (pmf):**
  - The pmf provides the probability of observing a certain number of events. Compute these probabilities using `dist.pmf(x_values)`.
- **Cumulative Distribution Function (CDF):**
  - The CDF provides the cumulative probability up to each number of events. Compute these cumulative probabilities using `dist.cdf(x_values)`.
- **Mean (Expected Value):**
  - The mean of a Poisson distribution is λ. Compute this using `dist.mean()`.
- **Variance:**
  - The variance of a Poisson distribution is also λ. Compute this using `dist.var()`.
- **Probability Generating Function (PGF):**
  - The PGF for a Poisson distribution is given by $G_X(t) = \exp(\lambda \cdot (\exp(t) - 1))$. Define a function `pgf(t, lambda_)` to compute PGF values for specific t values (e.g., 0 and 1).

6.4.8 Exercises

1. The number of patients arriving at a clinic follows a Poisson distribution with a mean of 3 patients per hour.
   (a) What is the probability that exactly 5 patients will arrive in an hour?
   (b) What is the probability that at most 2 patients will arrive in an hour?
   (c) What is the expected number of patients arriving in 3 hours?
2. A call center receives an average of 10 calls per hour. Assume the number of calls follows a Poisson distribution.
   (a) What is the probability that the call center receives exactly 12 calls in an hour?
   (b) Calculate the probability of receiving fewer than 5 calls in an hour.
   (c) Determine the probability of receiving more than 15 calls in an hour.
3. On a particular stretch of highway, the number of traffic accidents follows a Poisson distribution with an average rate of 3 accidents per month.

DRAFT

(a) What is the probability of exactly 2 accidents occurring in a month?

(b) Find the probability that there will be no accidents in a given month.

(c) Calculate the probability of having 4 or more accidents in a month.

4. A rare disease affects an average of 0.2 patients per 1000 individuals in a population. Assume the number of affected individuals follows a Poisson distribution.

   (a) What is the probability of finding exactly 1 patient with the disease in a sample of 1000 individuals?

   (b) Determine the probability of finding no patients with the disease in a sample of 1000 individuals.

   (c) Find the probability of discovering 2 or more patients with the disease in a sample of 1000 individuals.

5. An employee receives an average of 8 emails per day. Assume the number of emails follows a Poisson distribution.

   (a) What is the probability of receiving exactly 10 emails in a day?

   (b) Calculate the probability of receiving fewer than 6 emails in a day.

   (c) Determine the probability of receiving more than 12 emails in a day.

6. A retail store has an average of 20 customers arriving per hour. The number of customer arrivals follows a Poisson distribution.

   (a) What is the probability that exactly 25 customers will arrive in an hour?

   (b) Find the probability that fewer than 15 customers arrive in an hour.

   (c) Calculate the probability of having 30 or more customers in an hour.

7. A factory produces an average of 2 defective items per 1000 items produced. Assume the number of defective items follows a Poisson distribution.

   (a) What is the probability of finding exactly 3 defective items in a batch of 1000 items?

   (b) Determine the probability of finding no defective items in a batch of 1000 items.

   (c) Calculate the probability of finding 1 or more defective items in a batch of 1000 items.

8. In a laboratory experiment, rare events occur at an average rate of 0.4 events per hour. Assume these events follow a Poisson distribution.

   (a) What is the probability of observing exactly 2 events in an hour?

DRAFT

(b) Find the probability of observing no events in an hour.

(c) Determine the probability of observing more than 1 event in an hour.

9. A software program encounters an average of 5 errors per week. Assume the number of errors follows a Poisson distribution.

   (a) What is the probability of encountering exactly 7 errors in a week?

   (b) Calculate the probability of encountering fewer than 4 errors in a week.

   (c) Find the probability of encountering 8 or more errors in a week.

10. A library has an average of 12 book checkouts per day. The number of checkouts follows a Poisson distribution.

    (a) What is the probability of exactly 10 book checkouts in a day?

    (b) Determine the probability of having 15 or more book checkouts in a day.

    (c) Calculate the probability of having fewer than 8 book checkouts in a day.

11. In a small town, a rare disease has an average occurrence rate of 1 case per month. Assume the number of cases follows a Poisson distribution.

    (a) What is the probability of having exactly 2 cases of the disease in a month?

    (b) Find the probability of having no cases of the disease in a month.

    (c) Calculate the probability of having at least 1 case of the disease in a month.

12. In a certain industrial facility, accidents occur infrequently. It is known that the probability of an accident on any given day is 0.005 and accidents are independent of each other.

    (a) What is the probability that in any given period of 400 days there will be an accident on one day?

    (b) What is the probability that there are at most three days with an accident?

13. In a manufacturing process where glass products are made, defects or bubbles occur, occasionally rendering the piece undesirable for marketing. It is known that, on average, 1 in every 1000 of these items produced has one or more bubbles. What is the probability that a random sample of 8000 will yield fewer than 7 items possessing bubbles?

DRAFT

6.5 The Geometric and Negative Binomial Distributions

We have already explored two very important foundational discrete probability models: the **Binomial distribution**, which is applicable for modeling the number of successes in a fixed number of independent trials (such as counting heads in 10 coin tosses), and the **Poisson distribution**, which captures the count of rare, independent events over a fixed interval of time or space (such as incoming calls at a help desk). These distributions are essential tools in statistical modeling, but as with many theoretical models, the assumptions they rely on are often too restrictive for real-world data.

The Problem of Excess Variability

Suppose we are counting events, such as the number of customers arriving at a shop, the number of defects in manufactured items, or the number of bugs reported in a software system. If these events occur independently and at a constant average rate, the Poisson model is a natural choice. One of its defining properties is that the mean and variance of the distribution are equal.

However, this assumption often breaks down in real data. Consider a public health researcher tracking weekly flu cases across several cities. While the Poisson model may serve as a starting point, the actual data may reveal that the variance in flu counts is significantly larger than the mean. This is known as **overdispersion**, a common issue in count data.

Overdispersion suggests that events are not occurring as independently or uniformly as the Poisson model assumes. There may be unobserved heterogeneity such as differences in local healthcare access, variations in population density, or localized outbreaks that introduce additional variability. The Poisson model, with its strict equality of mean and variance, cannot accommodate such complexities.

Counting Trials Instead of Outcomes

Now, consider a different modeling question. The Binomial distribution answers the question: "How many successes will I observe in a fixed number of trials?" But what if we reverse the question: "How many trials (or failures) are required to achieve a fixed number of successes?"

This question leads us to two related distributions: the **Geometric distribution**, which models the number of failures before the *first* success, and the **Negative Binomial distribution**, which generalizes this idea to the case of *multiple* successes.

DRAFT

6.5.1 The Geometric Distribution

The Geometric distribution models the number of failures before the first success in a sequence of independent Bernoulli trials. Let X denote the number of failures before the first success, where each trial has success probability $p \in (0, 1)$. Then X follows a **Geometric distribution**, denoted

$$X \sim \text{Geom}(p),$$

with probability mass function:

$$P(X = k) = (1 - p)^k p, \quad \text{for } k = 0, 1, 2, \ldots$$

Definition: A random variable X is said to follow a Geometric distribution with parameter p, written as $X \sim \text{Geom}(p)$, if its probability mass function is given by

$$P(X = x) = (1 - p)^k p \quad \text{for } k = 0, 1, 2, 3, \ldots.$$

where,

- $p \in (0, 1)$ is the probability of success on each trial,
- X counts the number of failures before the first success.

For a geometric distribution with probability of success p, the cumulative distribution function (CDF) gives the probability that the first success occurs on or before the k-th trial. The CDF is defined as

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} (1 - p)^k \, p = p \sum_{k=0}^{x} (1 - p)^k$$

This is a **finite geometric series** of the form $\sum_{k=0}^{n} r^k = \dfrac{1 - r^{n+1}}{1 - r}$, with $r = (1 - p)$ and $n = x$. Thus we find,

$$\begin{aligned} F(x) &= p \cdot \frac{1 - (1 - p)^{x+1}}{1 - (1 - p)} = p \cdot \frac{1 - (1 - p)^{x+1}}{p} \\ &= \frac{p\big[1 - (1 - p)^{x+1}\big]}{p} = 1 - (1 - p)^{x+1}, \qquad x = 0, 1, 2, \ldots \end{aligned}$$

6.5.2 Expected Value and Variance of Geometric Distribution

Let $X \sim \text{Geometric}(p)$ with probability mass function

$$P(X = k) = (1 - p)^k \, p, \qquad k = 0, 1, 2, \ldots$$

For notational convenience, we write $q = 1 - p$ throughout.

DRAFT

Expected Value

The expected value of X is

$$E[X] = \sum_{k=0}^{\infty} k\, P(X = k) = \sum_{k=0}^{\infty} k\,(1-p)^k\, p = p \sum_{k=0}^{\infty} k\,(1-p)^k.$$

Since the $k = 0$ term contributes nothing, we write

$$E[X] = p \sum_{k=1}^{\infty} k\,(1-p)^k = p(1-p) \sum_{k=1}^{\infty} k\,(1-p)^{k-1}.$$

Using the series formula

$$\sum_{k=1}^{\infty} k\, r^{k-1} = \frac{1}{(1-r)^2}, \qquad |r| < 1,$$

with $r = 1 - p$, we obtain

$$E[X] = p(1-p) \cdot \frac{1}{(1-(1-p))^2} = p(1-p) \cdot \frac{1}{p^2} = \frac{1-p}{p}.$$

Thus the expected number of failures before the first success is $\dfrac{1-p}{p}$. Intuitively, the rarer the success (small p), the longer one expects to wait, which is reflected in the fact that $E[X]$ increases as p decreases.

Variance

The variance is given by

$$\mathrm{Var}(X) = E[X^2] - \big(E[X]\big)^2,$$

where

$$E[X^2] = \sum_{k=0}^{\infty} k^2\, P(X = k) = p \sum_{k=1}^{\infty} k^2\,(1-p)^k = p(1-p) \sum_{k=1}^{\infty} k^2\,(1-p)^{k-1}.$$

Using the series formula

$$\sum_{k=1}^{\infty} k^2\, r^{k-1} = \frac{1+r}{(1-r)^3}, \qquad |r| < 1,$$

with $r = 1 - p$, we have

$$E[X^2] = p(1-p) \cdot \frac{1+(1-p)}{(1-(1-p))^3} = p(1-p) \cdot \frac{2-p}{p^3} = \frac{(1-p)(2-p)}{p^2}.$$

DRAFT

Thus,

$$\begin{aligned}\mathrm{Var}(X) &= \frac{(1-p)(2-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} \\ &= \frac{(1-p)\big[(2-p)-(1-p)\big]}{p^2} = \frac{1-p}{p^2}.\end{aligned}$$

The standard deviation is thus $\mathrm{SD}(X) = \dfrac{\sqrt{1-p}}{p}$. As with the mean, both the variance and standard deviation grow without bound as $p \to 0$, capturing the increasing unpredictability of waiting times when successes are rare.

Problem 6.9. *A website has a 20% probability that a visitor makes a purchase on any given visit, independent of previous visits.*

(i) What is the probability that the first purchase occurs on the 4th visit?

(ii) What is the probability that the first purchase occurs on or before the 4th visit?

(iii) What is the expected number of visits before the first purchase?

(iv) What is the variance of the number of visits before the first purchase?

Solution:

Let $X \sim \text{Geometric}(p = 0.2)$, where X denotes the number of **visits without a purchase before the first purchase** is made, with PMF

$$P(X = k) = (1-p)^k\, p, \qquad k = 0, 1, 2, \ldots$$

The first purchase occurring on the 4th visit corresponds to exactly 3 failed visits beforehand, i.e. $X = 3$.

(i). The first purchase on the 4th visit means $X = 3$:

$$P(X = 3) = (1-p)^3 \cdot p = (0.8)^3 \cdot 0.2 = 0.512 \times 0.2 = 0.1024$$

There is a **10.24%** chance that the first purchase occurs exactly on the 4th visit, preceded by exactly 3 unsuccessful visits.

(ii). The first purchase on or before the 4th visit means at most 3 failed visits beforehand, i.e. $X \leq 3$. Using the CDF $F(x) = 1 - (1-p)^{x+1}$:

$$P(X \leq 3) = 1 - (1-p)^{3+1} = 1 - (0.8)^4 = 1 - 0.4096 = 0.5904$$

There is a **59.04%** chance that the first purchase is made within the first 4 visits.

DRAFT

(iii). The expected number of visits *before* the first purchase is

$$E[X] = \frac{1-p}{p} = \frac{0.8}{0.2} = 4$$

On average, a visitor makes 4 unsuccessful visits before completing a purchase, meaning the first purchase is expected on the 5th visit overall.

(iv). The variance of the number of visits before the first purchase is

$$\text{Var}(X) = \frac{1-p}{p^2} = \frac{0.8}{(0.2)^2} = \frac{0.8}{0.04} = 20$$

There is considerable variability around the mean of 4 failed visits, with a variance of 20. This reflects that while 4 failures is the average, the actual number of visits before a purchase can differ widely from customer to customer. □

Memoryless Property of the Geometric Distribution

Theorem 6.2. *(Memoryless Property of the Geometric Distribution) The geometric distribution with parameter p has the* memoryless property*, which means that for all integers* $m, n \geq 0$*,*

$$P(X > m + n \mid X > m) = P(X > n).$$

That is, given that the first success has not occurred by trial m*, the probability that it takes more than* n *additional trials is the same as the original probability that it takes more than* n *trials.*

Proof.

$$P(X > m + n \mid X > m) = \frac{P(X > m + n)}{P(X > m)}.$$

Since

$$P(X > k) = (1-p)^{k+1},$$

we have

$$\frac{P(X > m + n)}{P(X > m)} = \frac{(1-p)^{m+n+1}}{(1-p)^{m+1}} = (1-p)^{n+1} = P(X > n).$$

Hence, the geometric distribution is memoryless. □

Remark 6.5.1. *The geometric distribution is the only discrete distribution possessing this property. Its continuous counterpart is the exponential distribution.*

Problem 6.10. *A customer support center has observed that each call has a 30% chance of resolving a customer's issue. The outcome of each call is independent of previous ones. Given that the issue was not resolved in the first 2 calls, what is the probability that it will be resolved on the 5th call?*

DRAFT

Solution: Let X be the number of **failed calls before** the issue is resolved, so that $X \sim \text{Geometric}(p = 0.3)$ with PMF

$$P(X = k) = (1 - p)^k \, p, \qquad k = 0, 1, 2, \ldots$$

Resolution on the 5th call means exactly 4 failed calls preceded it, so we seek $P(X = 4 \mid X \geq 2)$. By the memoryless property of the geometric distribution,

$$P(X = 4 \mid X \geq 2) = P(X = 2)$$

since conditioning on at least 2 failures having already occurred is equivalent to restarting the process, with the remaining wait now requiring exactly 2 more failures. Therefore,

$$\begin{aligned} P(X = 4 \mid X \geq 2) &= P(X = 2) \\ &= (1 - 0.3)^2 \times 0.3 \\ &= (0.7)^2 \times 0.3 \\ &= 0.49 \times 0.3 \\ &= 0.147 \end{aligned}$$

The probability that the issue is resolved on the 5th call, given that it was not resolved in the first two, is **14.7%**. This result illustrates the memoryless property of the geometric distribution — the history of failed calls carries no information about future outcomes, so the process effectively restarts after each failure.

6.5.3 Applications of the Geometric Distribution

The geometric distribution is useful in modeling the number of trials until the first success. It finds various applications in data science, particularly in scenarios involving discrete waiting times, rare events, and user behavior modeling.

1. **Customer Behavior Modeling**
   - Used to model the number of user interactions (e.g., visits, ad impressions) before a conversion (e.g., a purchase).
   - Example: Probability that a user purchases on the 5^{th} website visit.
2. **Clickstream and Funnel Analysis**
   - Models the number of web page visits before a key action (e.g., signing up or clicking "Buy").
   - Useful for analyzing conversion rates in A/B testing and digital marketing.

DRAFT

3. **Quality Control and Reliability Engineering**
    - Models the number of items inspected before finding a defective product.
    - Applied in manufacturing and operational processes.
4. **Simulation and Synthetic Data Generation**
    - Simulates trial-until-success processes, useful in experimental design and bootstrapped datasets.
5. **Fraud Detection**
    - Models the number of normal transactions before detecting a fraudulent one.
    - Helps identify and respond to rare events.
6. **Telecommunications and Networking**
    - Models the number of retries before a successful packet transmission.
    - Useful in error modeling in network protocols.
7. **Natural Language Processing (NLP)**
    - Occasionally used to model the number of words or tokens until a specific event (e.g., a keyword) occurs.
8. **Survival Analysis (Discrete-Time)**
    - Acts as the discrete-time analog of the exponential distribution.
    - Models the number of periods before an event such as dropout or failure occurs.

These applications take advantage of the geometric distribution's memoryless property and simple parameterization, making it suitable for probabilistic modeling in a variety of data science contexts.

6.5.4 The Negative Binomial Distribution

The Negative Binomial distribution generalizes the Geometric by modeling the number of failures before achieving a total of r successes. Let the random variable X denote the **number of failures before the r-th success**, where $r \in \mathbb{N}$ and each trial has success probability $p \in (0, 1)$. Then X is said to follow a **Negative Binomial distribution**, denoted

$$X \sim \text{NegBin}(r, p),$$

DRAFT

with probability mass function:

$$P(X = k) = \binom{k + r - 1}{k}(1 - p)^k p^r, \quad \text{for } x = 0, 1, 2, \ldots \tag{6.1}$$

where,

- $r > 0$: number of required successes
- $p \in (0, 1)$: probability of success on each trial

The mean and variance of the Negative Binomial distribution are

$$E(X) = \frac{r(1-p)}{p} \qquad \text{and} \qquad \text{Var}(X) = \frac{r(1-p)}{p^2},$$

respectively. Since $0 < p < 1$, the variance is always greater than the mean, a property known as **overdispersion**.

Definition: A random variable X is said to follow a Negative Binomial distribution with parameters r and p, written as $X \sim \text{NegBin}(r, p)$, if its probability mass function is given by

$$P(X = k) = \binom{k + r - 1}{k}(1 - p)^k p^r \quad \text{for } k = 0, 1, 2, 3, \ldots.$$

where,

- $r \in \mathbb{N}$ is the number of successes,
- $p \in (0, 1)$ is the success probability on each trial.

The Negative Binomial distribution is suitable for modeling overdispersed count data.

Problem 6.11. *A salesperson has a 40% chance of closing a deal on any given sales call. Each call is independent of the others. What is the probability that the salesperson will close their **3rd deal on the 7th call**? Also, compute the expected number of calls and the **variance** required to make 3 successful sales.*

Solution: Let $X \sim \text{NegBin}(r = 3, p = 0.4)$, where X is the number of failures before the 3rd success.

(i) Probability of 4 failures before the 3rd success:

$$P(X = 4) = \binom{4 + 3 - 1}{4}(1 - 0.4)^4(0.4)^3 = \binom{6}{4}(0.6)^4(0.4)^3$$
$$= 15 \cdot 0.1296 \cdot 0.064 = 0.1244$$

There is a 12.44% chance of experiencing exactly 4 failures before resolving the 3rd issue.

(ii) Expected number of failures:

$$E[X] = \frac{r(1-p)}{p} = \frac{3 \cdot 0.6}{0.4} = 4.5$$

(iii) Variance:

$$\text{Var}(X) = \frac{r(1-p)}{p^2} = \frac{3 \cdot 0.6}{0.16} = 11.25$$

On average, 4.5 failures are expected before achieving 3 successes, with a variance of 11.25.

□

Problem 6.12. *When a fisherman catches a fish, it is a young one with a probability of 0.23 and it is returned to the water. On the other hand, an adult fish is kept to be eaten later.*

(a) What is the expected number of fish caught by the fisherman before an adult fish is caught?

(b) What is the probability that the fifth fish caught is the first young fish?

(c) Suppose that the fisherman wants three fish to eat for lunch.

(i) What is the probability that the first time the fisherman can stop for lunch is immediately after the sixth fish has been caught?

(ii) If the fisherman catches eight fish, what is the probability that there are sufficient fish for lunch?

Solution: A fish caught is young with probability $p = 0.23$ (and is returned), and adult with probability $q = 1 - p = 0.77$ (and is kept).

(a) **Expected number of fish caught before an adult fish is caught:**

Let X be the number of fish caught until the first adult fish is caught. Then $X \sim \text{Geometric}(q = 0.77)$. The expected value is:

$$E[X] = \frac{1}{q} = \frac{1}{0.77} \approx 1.30$$

So, on average, the fisherman will catch about 1.30 fish before catching an adult.

(b) **Probability that the fifth fish caught is the first young fish:** Let $Y \sim \text{Geometric}(p = 0.23)$, where Y is the number of trials until the first young fish is caught.

$$P(Y = k) = (1-p)^{k-1} \cdot p = (0.77)^{k-1} \cdot 0.23, \quad \text{for } k = 1, 2, 3, \ldots$$

DRAFT

To compute the probability that the first young fish is caught on the 5^{th} trial, plug in $k = 5$:

$$P(Y = 5) = (0.77)^4 \cdot 0.23 \approx 0.3515 \cdot 0.23 \approx 0.0808$$

(c) **The fisherman wants 3 adult fish for lunch.**

(i). **Probability that the first time he can stop for lunch is after catching the sixth fish:**

Let Z be the trial on which the 3^{rd} adult fish is caught. Then $Z \sim \text{NegativeBinomial}(r = 3, p = 0.77)$, and we want:

$$P(Z = 6) = \binom{5}{2} \cdot (0.77)^3 \cdot (0.23)^3 = 10 \cdot 0.457 \cdot 0.0122 \approx 0.0558$$

(ii). **Probability that there are at least 3 adult fish among 8 catches:**

Let W be the number of adult fish among 8 catches. Then $W \sim \text{Binomial}(n = 8, p = 0.77)$. Wa want the probability at least 3 adult fish among 8 catches. Then

$$P(W \geq 3) = 1 - P(W < 3) = 1 - [P(W = 0) + P(W = 1) + P(W = 2)]$$

$$P(W = 0) = \binom{8}{0}(0.77)^0(0.23)^8 \approx 7.83 \times 10^{-6}$$

$$P(W = 1) = \binom{8}{1}(0.77)^1(0.23)^7 \approx 0.00021$$

$$P(W = 2) = \binom{8}{2}(0.77)^2(0.23)^6 \approx 0.00246$$

$$\text{Hence, } P(W < 3) \approx 0.0027 \quad \Rightarrow \quad P(W \geq 3) \approx 1 - 0.0027 = 0.9973$$

So, the fisherman is almost certain to have enough adult fish for lunch after catching 8 fish.

```python
from scipy.stats import geom, nbinom, binom

# Given probabilities
p_young = 0.23 # Probability of young fish (returned)
p_adult = 1 - p_young # Probability of adult fish (kept)

print("(a) Expected number of fish until first adult")
# Geometric distribution: X ~ Geometric(p_adult)
expected_X = 1 / p_adult
print(f"Expected number of fish: {expected_X:.2f}")

```

DRAFT

```
print("\n (b) Probability that 5th fish is the first young fish")
# Y ~ Geometric(p_young), since we want first young fish
k = 5
P_Y_5 = geom.pmf(k, p_young)  # Geometric pmf at k
print(f"P(Y = 5) = {P_Y_5:.4f}")

print("\n (c)(i) Probability third adult is caught on 6th trial")
# Z ~ NegativeBinomial(r=3, p=p_adult)
# This is P(Z = 6), i.e., 2 adult fish in first 5, 3rd on 6th
r = 3
n = 6
P_Z_6 = nbinom.pmf(n - r, r, p_adult)  # nbinom.pmf(k_failures, r_successes, p)
print(f"P(Z = 6) = {P_Z_6:.4f}")

print("\n (c)(ii) Probability at least 3 adult fish in 8 catches")
# W ~ Binomial(n=8, p=p_adult)
n_total = 8
k_min = 3
P_W_ge_3 = 1 - binom.cdf(k_min - 1, n_total, p_adult)
print(f"P(W >= 3) = {P_W_ge_3:.4f}")
```

6.5.5 Why Is It Called the Negative Binomial Distribution?

This distribution is called *negative binomial* because its probability mass function involves coefficients from the negative binomial series expansion, and it generalizes the binomial distribution to a setting where the number of trials is random and the number of successes is fixed. The name comes from the *negative binomial series*:

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \binom{k+r-1}{k} x^k, \quad \text{for } |x| < 1,\ r > 0$$

This series is similar to the binomial expansion, but it allows for negative or non-integer exponents.

The "negative" in **negative binomial distribution** precisely refers to its connection with the **negative binomial series** (or binomial series with a negative exponent). In density given in (6.1), the crucial part is the binomial coefficient: $\binom{k+r-1}{k}$. A key property of binomial coefficients states that for any

DRAFT

real number α and non-negative integer k is

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\dots(\alpha-k+1)}{k!}$$

Now, consider the case where $\alpha = -r$ (where r is a positive integer).

$$\binom{-r}{k} = \frac{(-r)(-r-1)\dots(-r-k+1)}{k!}$$

We can factor out $(-1)^k$ from the numerator:

$$\binom{-r}{k} = \frac{(-1)^k r(r+1)\dots(r+k-1)}{k!}$$

The term $r(r+1)\dots(r+k-1)$ can be written as $\frac{(r+k-1)!}{(r-1)!}$. So,

$$\binom{-r}{k} = (-1)^k \frac{(r+k-1)!}{k!(r-1)!} = (-1)^k \binom{r+k-1}{k}$$

From this, we can see that:

$$\binom{r+k-1}{k} = (-1)^k \binom{-r}{k}$$

Substituting this back into the pmf of the negative binomial distribution:

$$P(X=k) = (-1)^k \binom{-r}{k} p^r (1-p)^k$$

This form directly links the pmf to the terms in the **negative binomial series expansion** of $(1-x)^{-r}$:

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \binom{-r}{k} (-x)^k = \sum_{k=0}^{\infty} \binom{-r}{k} (-1)^k x^k$$

If we substitute $x = (1-p)$ into this series, the terms of the expansion are proportional to the probabilities of the negative binomial distribution. The sum of these probabilities over all possible k equals 1, just as the sum of the series when $x = (1-p)$ equals $(1-(1-p))^{-r} = (p)^{-r}$.

Thus, the name "negative binomial" arises from the mathematical identity that allows the probability mass function to be expressed using a binomial coefficient with a negative upper index, connecting it to the negative binomial series.

DRAFT

Different Representations of the Negative Binomial Distribution

The Negative Binomial distribution can be parameterized in several equivalent forms, depending on the interpretation of the random variable. One commonly used form is given in Equation (6.1). Let $Y = X + r$. Then Y represents the **total number of trials** required to achieve the rth success. The probability mass function of Y is

$$P(Y = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad \text{for } n = r, r+1, \ldots$$

This is a reparameterization of the density given in Equation (6.1).

Another, less common, parameterization considers the number of **successes before the rth failure**, treating a "failure" as the event of interest. In this case, the probability mass function for the number of successes before the rth failure is

$$P(Z = k) = \binom{k+r-1}{k} p^k (1-p)^r, \quad \text{for } k = 0, 1, 2, \ldots$$

where Z is the number of successes observed before the rth failure. All reparameterized forms of the density are summarized in the following table.

Random Variable	**pdf**	**Support**
Failures before rth success	$\binom{k+r-1}{k}(1-p)^k p^r$	$k = 0, 1, \ldots$
Trials until rth success	$\binom{n-1}{r-1} p^r (1-p)^{n-r}$	$n = r, r+1, \ldots$
Successes before rth failure	$\binom{k+r-1}{k} p^k (1-p)^r$	$k = 0, 1, \ldots$

6.5.6 Relation to Geometric

The Geometric distribution can be viewed as a special case of the Negative Binomial distribution where the number of required successes r is 1. While the Geometric counts trials until the first success, the Negative Binomial generalizes this to count trials until the r-th success. Both distributions share a similar functional form, but the Negative Binomial can model more complex scenarios requiring multiple successes.

Importance of the Negative Binomial Distribution

The Negative Binomial distribution is important in both theoretical and applied statistics for several key reasons. Some of them are mentioned in the following:

DRAFT

- **Modeling Overdispersed Data:** It is often used to model count data with overdispersion, where the variance exceeds the mean, unlike the Poisson distribution.
- **Generalization of the Geometric Distribution:** It extends the geometric distribution by allowing the count of failures before r successes, providing more modeling flexibility.
- **Applications in Real Life:** Common in fields such as epidemiology, finance, and insurance where event probabilities vary and simple models like binomial or Poisson are insufficient.
- **Used in Bayesian and Compound Models:** It emerges naturally as a Poisson-Gamma mixture model, making it useful in Bayesian inference and hierarchical modeling.

Whether you are analyzing noisy count data or sequential trial outcomes, the Negative Binomial distribution is a powerful tool for modern applied statistics.

6.5.7 Relationship Between Binomial and Negative Binomial Distributions

The Binomial and Negative Binomial distributions are both discrete probability distributions that model the outcomes of Bernoulli trials (experiments with two possible outcomes: success or failure). However, they differ fundamentally in what is considered fixed and what is random. The Binomial distribution models the number of *successes* in a fixed number of independent trials, while the Negative Binomial distribution models the number of *failures* (or total trials) required to achieve a fixed number of successes. The following table summarizes the key differences between them:

Table 6.1: Comparison between Binomial and Negative Binomial Distributions

Feature	**Binomial Distribution**	**Negative Binomial Distribution**
Fixed Value	Number of trials n	Number of successes r
Random Variable	Number of successes in n trials	Number of failures (or trials) until r successes
Question Answered	How many successes in a fixed number of tries?	How many failures (or tries) to get fixed successes?
Example	Getting 3 heads in 5 coin flips	Getting 3 heads in X coin flips (until 3rd head)

DRAFT

6.5.8 Probability Generating Function

The probability generating function of Negative Binomial distribution is given in the Theorem 6.3.

Theorem 6.3 (Probability Generating Function)**.** *Let $X \sim NegBin(r, p)$ be a Negative Binomial random variable representing the number of failures before the r-th success, where $r > 0$ and $0 < p < 1$. Then the probability generating function (PGF) of X is given by*

$$G_X(s) = E[s^X] = \left(\frac{p}{1-(1-p)s}\right)^r, \quad \text{for } |(1-p)s| < 1.$$

Proof. By definition, the probability generating function of a discrete random variable X is

$$G_X(s) = E[s^X] = \sum_{x=0}^{\infty} P(X = x)s^x.$$

The probability mass function of the Negative Binomial distribution is

$$P(X = x) = \binom{x+r-1}{x}(1-p)^x p^r, \quad x = 0, 1, 2, \ldots$$

Substituting into the PGF definition, we get

$$G_X(s) = \sum_{x=0}^{\infty} \binom{x+r-1}{x}(1-p)^x p^r s^x = p^r \sum_{x=0}^{\infty} \binom{x+r-1}{x} ((1-p)s)^x.$$

Recall the identity (valid for $|a| < 1$):

$$\sum_{x=0}^{\infty} \binom{x+r-1}{x} a^x = (1-a)^{-r}.$$

Letting $a = (1-p)s$, we obtain

$$G_X(s) = p^r \cdot (1-(1-p)s)^{-r} = \left(\frac{p}{1-(1-p)s}\right)^r.$$

This expression is valid for $|(1-p)s| < 1$, ensuring the convergence of the series. □

6.5.9 Moment Generating Function

The moment generating function of Negative Binomial distribution is given in the Theorem 6.4.

DRAFT

Theorem 6.4 (Moment Generating Function)**.** *Let $X \sim NegBin(r, p)$ be a Negative Binomial random variable representing the number of failures before the r-th success, where $r > 0$ and $0 < p < 1$. Then the moment generating function of X is given by*

$$M_X(t) = E[e^{tX}] = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad \text{for } (1-p)e^t < 1.$$

Proof. By definition, the moment generating function of a random variable X is

$$M_X(t) = E[e^{tX}] = \sum_{x=0}^{\infty} P(X = x)e^{tx}.$$

The pmf of the Negative Binomial distribution is

$$P(X = x) = \binom{x+r-1}{x}(1-p)^x p^r, \quad x = 0, 1, 2, \ldots$$

Substituting into the definition,

$$M_X(t) = \sum_{x=0}^{\infty} \binom{x+r-1}{x}(1-p)^x p^r e^{tx} = p^r \sum_{x=0}^{\infty} \binom{x+r-1}{x} \left((1-p)e^t\right)^x.$$

Using the identity

$$\sum_{x=0}^{\infty} \binom{x+r-1}{x} a^x = (1-a)^{-r}, \quad \text{for } |a| < 1,$$

let $a = (1-p)e^t$. Then,

$$M_X(t) = p^r \cdot \left(1-(1-p)e^t\right)^{-r} = \left(\frac{p}{1-(1-p)e^t}\right)^r.$$

This result is valid for $(1-p)e^t < 1$. □

Theorem 6.5. *Let $X_1, X_2, \ldots, X_r$ be independent and identically distributed geometric random variables with success probability p, i.e., for each i,*

$$P(X_i = x) = (1-p)^x p, \quad x = 0, 1, 2, \ldots$$

Then the sum

$$Y = \sum_{i=1}^{r} X_i$$

follows a negative binomial distribution with parameters r and p. That is,

$$P(Y = k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad k = 0, 1, 2, \ldots$$

DRAFT

Proof. We will prove this theorem using moment generating functions (MGFs). The MGF uniquely determines a probability distribution.

The probability mass function of X_i is $P(X_i = x) = (1-p)^x p$ for $x = 0, 1, 2, \ldots$. The moment generating function $M_{X_i}(t)$ is defined as:

$$\begin{aligned} M_{X_i}(t) &= E[e^{tX_i}] \\ &= \sum_{x=0}^{\infty} e^{tx} P(X_i = x) \\ &= \sum_{x=0}^{\infty} e^{tx} (1-p)^x p \\ &= p \sum_{x=0}^{\infty} (e^t(1-p))^x \end{aligned}$$

This is a geometric series with first term 1 (when $x = 0$) and common ratio $e^t(1-p)$. For the series to converge, we must have $|e^t(1-p)| < 1$. The sum of an infinite geometric series $1 + a + a^2 + \ldots$ is $\frac{1}{1-a}$ for $|a| < 1$. So,

$$M_{X_i}(t) = p \times \frac{1}{1 - e^t(1-p)} = \frac{p}{1 - (1-p)e^t}$$

Since $X_1, X_2, \ldots, X_r$ are independent, the MGF of their sum is the product of their individual MGFs:

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t(X_1 + X_2 + \cdots + X_r)}] \\ &= E[e^{tX_1} e^{tX_2} \ldots e^{tX_r}] \end{aligned}$$

Due to independence, $M_Y(t) = E[e^{tX_1}]E[e^{tX_2}] \ldots E[e^{tX_r}]$. Since the X_i are identically distributed, $M_{X_i}(t)$ is the same for all i.

$$M_Y(t) = \left(\frac{p}{1 - (1-p)e^t} \right)^r$$

Using Theorem 6.4, we have

$$M_{NB}(t) = p^r (1 - (1-p)e^t)^{-r} = \frac{p^r}{(1 - (1-p)e^t)^r} = \left(\frac{p}{1 - (1-p)e^t} \right)^r$$

We observe that the MGF of $Y = \sum_{i=1}^{r} X_i$ is identical to the MGF of a negative binomial distribution with parameters r and p. Since the MGF uniquely determines the distribution, we conclude that Y follows a negative binomial distribution with parameters r and p.

□

DRAFT

6.5.10 Mean and Variance of the Negative Binomial Distribution

Let $Y \sim \text{NegBin}(r, p)$ be the number of failures before the r-th success in independent Bernoulli trials with success probability p. Equivalently, Y can be expressed as a sum of r i.i.d. geometric random variables:

$$Y = X_1 + X_2 + \cdots + X_r,$$

where each $X_i \sim \text{Geom}(p)$ (with support $\{0, 1, 2, \dots\}$) and

$$P(X_i = k) = (1-p)^k p, \quad k = 0, 1, 2, \dots$$

For a geometric random variable X_i,

$$E[X_i] = \sum_{k=0}^{\infty} k(1-p)^k p.$$

Using the formula for the expectation of a geometric distribution starting at 0, we have

$$E[X_i] = \frac{1-p}{p}.$$

Similarly, the variance of X_i is known to be

$$\text{Var}(X_i) = \frac{1-p}{p^2}.$$

Mean and variance of Y
Since Y is the sum of r i.i.d. X_i's, by linearity of expectation:

$$E[Y] = \sum_{i=1}^{r} E[X_i] = r \cdot \frac{1-p}{p}.$$

Because the X_i's are independent,

$$\text{Var}(Y) = \sum_{i=1}^{r} \text{Var}(X_i) = r \cdot \frac{1-p}{p^2}.$$

□

Alternatively, let $X \sim \text{NegBin}(r, p)$, where X is the number of failures before observing r successes, and p is the probability of success on each trial. Using Theorem 6.5, the moment generating function of X is given by:

$$M_X(t) = \left(\frac{p}{1-(1-p)e^t} \right)^r, \quad \text{for } t < -\ln(1-p)$$

DRAFT

Mean: To find the mean, we differentiate the MGF and evaluate at $t = 0$:

$$E[X] = M'_X(0) = r \cdot \frac{1-p}{p}$$

Variance: To find the variance, we compute the second derivative of the MGF:

$$\text{Var}(X) = M''_X(0) - (M'_X(0))^2 = \frac{r(1-p)}{p^2}$$

6.5.11 When to Use the Negative Binomial Distribution

The Negative Binomial distribution should be considered whenever:

- Count data exhibit greater variability than the Poisson model allows.
- The data include many zeros and a long right tail.
- There is reason to believe that the event rate differs across units or conditions.

This distribution is widely used in fields such as epidemiology, ecology, marketing, and insurance, where overdispersed counts are common.

6.5.12 Exercises

1. A fair coin is tossed repeatedly. What is the probability that the first head appears on the 5^{th} toss?

2. In a quality test, the probability that a product is defective is 0.05. What is the probability that the first defective product is found within the first 10 items tested?

3. A basketball player has a 40% chance of making any free throw. Calculate the probability that the player misses the first 4 shots and makes the 5^{th}.

4. A software company tracks the number of customer support calls needed before a user's issue is resolved. Suppose the probability that an issue is resolved on any given call is 0.25, and each call is independent of the others.

   (a) What is the probability that the issue is resolved on the 3rd call?

   (b) What is the probability that it takes at most 3 calls to resolve the issue?

   (c) What is the expected number of calls required to resolve the issue?

   (d) What is the variance of the number of calls?

DRAFT

5. A basketball player has a probability of 0.6 of making a free throw. Find the probability that she misses exactly 3 shots before making her 5^{th} successful shot.

6. Let $X \sim \text{NegBin}(3, 0.4)$. Compute the probability that there are at most 4 failures before the third success, i.e., $P(X \leq 4)$.

7. If $X \sim \text{NegBin}(4, 0.5)$, compute:

   (a) the mean number of failures,

   (b) the variance of the number of failures.

8. Compare $X_1 \sim \text{NegBin}(2, 0.3)$ and $X_2 \sim \text{NegBin}(4, 0.6)$. Which has the higher mean number of failures, and which is more variable?

9. A customer service agent closes a sale with probability 0.2. Find the probability that she gets her 3^{rd} sale after exactly 7 rejections.

10. Each item has a 10% chance of being defective. Find the probability that 2 defectives are found after 8 non-defective items.

11. A quality inspector wants to find 5 defective parts. If each part has a 0.1 chance of being defective, find the probability she inspects exactly 30 parts to find the 5 defectives.

12. A worker fails 12 times before achieving 4 successes. Estimate the success probability p using the method of moments.

13. A gamer has a 30% chance of winning a match. What is the probability it takes her between 5 and 8 losses (inclusive) to win her 4^{th} game?

14. A machine produces items, and it is known that a certain percentage of the items produced are defective. The probability of an item being defective is 0.05. Assume that the production of each item is an independent event.

    (a) What is the expected number of items produced until the first defective item is found?

    (b) What is the probability that the third item produced is the first one that is defective?

    (c) Suppose a quality control inspector checks a batch of 10 items.

        (i) What is the probability that the inspector finds exactly 2 defective items in the batch?

        (ii) What is the probability that the inspector finds at least 1 defective item in the batch?

15. n a particular online game, a player has a 0.40 probability of winning a round. Each round is independent of the others.

DRAFT

(a) What is the expected number of rounds a player needs to play to achieve their first win?

(b) What is the probability that the player's fourth round is their first win?

(c) If a player plays 7 rounds,

(i) What is the probability that they win exactly 3 of these rounds?

(ii) What is the probability that they win more than 5 rounds?

6.6 Discrete Uniform Distribution

There are two types of uniform distributions: discrete uniform distribution and continuous uniform distribution. The **discrete uniform distribution** plays a crucial role in various domains, particularly in statistics and data science. At its core, the discrete uniform distribution describes a situation where a finite number of outcomes are equally likely to occur. This property makes it an essential model for understanding fundamental concepts in probability theory.

In the realm of **simulations**, particularly **Monte Carlo simulations**, the discrete uniform distribution is employed to generate random inputs. This is vital for modeling complex systems and assessing the behavior of different scenarios. Similarly, in A/B testing, where subjects are randomly assigned to different groups, the assumption of a uniform distribution ensures that the groups are comparable, thereby enhancing the validity of the test results.

In a discrete uniform distribution, a finite number of outcomes are possible, and each outcome has an equal probability of occurring. For example, if you roll a fair six-sided die, the probability of each face showing up (1 through 6) is the same

$$P(X = x) = \frac{1}{6} \quad ; \quad x = 1, 2, \ldots, 6.$$

Definition: A random variable X is said to follow a **discrete uniform distribution** on the set of integers $\{x_1, x_2, \ldots, x_n\}$, denoted $X \sim U(x_1, x_n)$, if its pmf is given by

$$P(X = x) = \frac{1}{n} \quad ; \quad x \in \{x_1, x_2, \ldots, x_n\},$$

where n is the total number of outcomes.

If the lower bound of X is zero and the upper bound is n then the pmf is

DRAFT

given by

$$P(X = x) = \frac{1}{n+1} \quad ; \quad x = 0, 1, 2, \ldots, n.$$

The discrete uniform distribution finds applications in games of chance. Understanding the probabilities associated with games like lotteries or card games helps players develop strategies and assess risks. Beyond these practical applications, the discrete uniform distribution is foundational for other probability distributions. Many advanced statistical methods and techniques build upon the principles established by the uniform distribution, making it a key component of statistical inference.

6.6.1 Mean and Variance of Discrete Uniform Distribution

Let $X \sim U(x_1, x_2, \ldots, x_n)$ be a discrete uniform distribution over the n distinct values $\{x_1, x_2, \ldots, x_n\}$, where all values are equally likely. Then the probability mass function (pmf) is:

$$P(X = x_i) = \frac{1}{n}, \quad \text{for } i = 1, 2, \ldots, n.$$

Expected Value: The expected value of X is:

$$E[X] = \sum_{i=1}^{n} x_i \cdot P(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

If X takes values in the set $\{1, 2, \ldots, n\}$ with equal probability, then the expectation is

$$E[X] = \frac{1}{n} \sum_{k=1}^{n} k = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

Variance: The variance of X is defined as:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

First compute $E[X^2]$:

$$E[X^2] = \sum_{i=1}^{n} x_i^2 \cdot P(X = x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Now compute the variance:

$$\mathrm{Var}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2$$

DRAFT

If X takes values in the set $\{1, 2, \ldots, n\}$ with equal probability, then

$$E[X^2] = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

Hence, the variance,

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

Simplify,

$$\mathrm{Var}(X) = \frac{(n+1)}{6}\left(2n + 1 - \frac{3(n+1)}{2}\right) = \frac{n^2-1}{12}$$

Using the pmf $P(X = x) = \frac{1}{n+1}$, for $x = 0, 1, 2, \ldots, n$. the mean and variance are given in Theorem 6.6.

Theorem 6.6. *Let X be a random variable that follows a discrete uniform distribution on the set $\{0, 1, 2, \ldots, n\}$. Then the mean μ and variance σ^2 of X are given by*

$$\mu = \frac{n}{2} \quad \textit{and} \quad \sigma^2 = \frac{n^2-1}{12}.$$

Proof. **Mean:** The mean μ of X is calculated as follows

$$\mu = E[X] = \sum_{x=0}^{n} x \cdot P(X = x) = \sum_{x=0}^{n} x \cdot \frac{1}{n+1} = \frac{1}{n+1}\sum_{x=0}^{n} x.$$

Using the formula for the sum of the first n integers, we have

$$\sum_{x=0}^{n} x = \frac{n(n+1)}{2}.$$

Substituting this back into the equation for μ

$$\mu = \frac{1}{n+1} \cdot \frac{n(n+1)}{2} = \frac{n}{2}.$$

Variance:

The variance σ^2 is calculated using the formula

$$\sigma^2 = E[X^2] - (E[X])^2.$$

First, we compute $E[X^2]$

$$E[X^2] = \sum_{x=0}^{n} x^2 \cdot P(X = x) = \sum_{x=0}^{n} x^2 \cdot \frac{1}{n+1}.$$

DRAFT

This simplifies to

$$E[X^2] = \frac{1}{n+1}\sum_{x=0}^{n} x^2.$$

Using the formula for the sum of the squares of the first n integers, we have

$$\sum_{x=0}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}.$$

Substituting this into the equation for $E[X^2]$:

$$E[X^2] = \frac{1}{n+1} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(2n+1)}{6}.$$

Now, we can compute the variance

$$\sigma^2 = E[X^2] - (E[X])^2 = \frac{n(2n+1)}{6} - \left(\frac{n}{2}\right)^2.$$

Thus,

$$\sigma^2 = \frac{2n(2n+1)}{12} - \frac{3n^2}{12} = \frac{4n^2+2n-3n^2}{12} = \frac{n^2+2n}{12} = \frac{n^2-1}{12}.$$

□

Problem 6.13. *A company has a database of 100 customer IDs ranging from 1 to 100. The marketing team wants to select a random sample of 10 customer IDs to send a promotional offer. Each customer ID should have an equal chance of being selected.*

(a). What is the probability of selecting any specific customer ID?

(b). If the marketing team selects 10 IDs, what is the expected number of times a specific customer ID (e.g., ID 5) will be included in the sample?

Solution

(a). Probability of Selecting Any Specific Customer ID

Since the customer IDs are uniformly distributed, the probability mass function (pmf) is

$$P(X = x) = \frac{1}{n} \quad ; \quad x = 1, 2, \ldots, 100.$$

Here, $n = 100$. Thus, the probability of selecting any specific customer ID (say ID 5) is

$$P(X = 5) = \frac{1}{100} = 0.01.$$

DRAFT

(b). Expected Number of Times a Specific Customer ID Will Be Included

Let

$$X = \begin{cases} 1, & \text{if ID 5 appears in the sample,} \\ 0, & \text{otherwise.} \end{cases}$$

Since an ID can appear at most once in a sample drawn **without replacement**,

$$E[X] = P(X = 1) = P(\text{ID 5 is selected}) = 0.10.$$

Thus, the expected number of times ID 5 appears in the sample is 0.10. This means, on average, out of many repeated samples of size 10, ID 5 will appear 10% of the time

Problem 6.14. *A school is organizing a raffle with 50 tickets numbered from 1 to 50. Each ticket has an equal chance of being drawn.*

(a). What is the probability of drawing any specific ticket number?

(b). If the raffle will draw 5 tickets, what is the expected number of times a specific ticket (e.g., ticket number 10) will be drawn?

Solution

(a). Probability of Drawing Any Specific Ticket Number

Since the tickets are uniformly distributed, the probability mass function (pmf) is given by

$$P(X = x) = \frac{1}{n} \quad ; \quad x = 1, 2, \ldots, 50.$$

Here, $n = 50$. Thus, the probability of drawing any specific ticket number (e.g., ticket number 10) is:

$$P(X = 10) = \frac{1}{50} = 0.02.$$

(b). Expected Number of Times a Specific Ticket Will Be Drawn

The raffle draws 5 tickets *without replacement*. A specific ticket can appear at most once.

Let

$$X = \begin{cases} 1, & \text{if ticket 10 is drawn among the 5 tickets,} \\ 0, & \text{otherwise.} \end{cases}$$

The probability that ticket 10 is included in the set of 5 tickets is

$$P(X = 1) = \frac{5}{50} = 0.10.$$

Thus, the expected number of times ticket 10 appears in the raffle draw is

$$E[X] = P(X = 1) = 0.10.$$

DRAFT

6.6.2 Python Code Discrete Uniform Distribution

```
import numpy as np

# Define the Discrete Uniform parameters
a_discrete = 1
b_discrete = 10

def pmf_discrete(x, a, b):
  if a <= x <= b:
    return 1 / (b - a + 1)
  else:
    return 0

def cdf_discrete(x, a, b):
  if x < a:
    return 0
  elif a <= x <= b:
    return (x - a + 1) / (b - a + 1)
  else:
    return 1

# PMF values
x_discrete_values = np.arange(a_discrete, b_discrete + 1)
pmf_values_discrete = [pmf_discrete(x, a_discrete,
  b_discrete) for x in x_discrete_values]
print("PMF values for x from {} to {}:".format(a_discrete,
   b_discrete), pmf_values_discrete)

# CDF values
cdf_values_discrete = [cdf_discrete(x, a_discrete,
  b_discrete) for x in x_discrete_values]
print("CDF values for x from {} to {}:".format(a_discrete,
   b_discrete), cdf_values_discrete)

# Mean (Expected Value)
mean_discrete = (a_discrete + b_discrete) / 2
print("Mean (Expected Value):", mean_discrete)

# Variance
variance_discrete = ((b_discrete - a_discrete + 1) ** 2 -
  1) / 12
print("Variance:", variance_discrete)

# Standard Deviation
std_dev_discrete = np.sqrt(variance_discrete)
print("Standard Deviation:", std_dev_discrete)
```

6.6.3 Exercises

1. A box contains 20 different colored balls, numbered from 1 to 20. If a ball is drawn at random, what is the probability of drawing ball number 15?

2. A survey is conducted with 30 participants, each assigned a unique ID from 1 to 30. If 5 IDs are randomly selected for a follow-up interview, what is the expected number of times a specific ID (e.g., ID 12) will be selected?

3. A raffle has 100 tickets numbered from 1 to 100. If the raffle draws 10 tickets, what is the probability that ticket number 25 will be drawn at least once?

4. In a game, a player rolls a fair six-sided die. What is the probability of rolling a specific number, say 4? Additionally, if the player rolls the die 10 times, what is the expected number of times the number 4 will appear?

5. A classroom has 15 students, each assigned a number from 1 to 15. If the teacher randomly selects 3 students for a project, what is the probability that student number 7 is selected?

6. A bag contains 50 different coupons numbered from 1 to 50. If 5 coupons are drawn randomly, what is the expected number of times coupon number 30 will be drawn?

7. A committee consists of 12 members, each assigned a number from 1 to 12. If 4 members are randomly chosen to form a subcommittee, what is the probability that member number 6 is included in the selection?

8. A local lottery involves selecting 6 numbers from a set of 1 to 49. What is the probability that the number 7 is chosen in a single drawing? If you play the lottery 10 times, what is the expected number of times you will have number 7 in your selected numbers?

9. In a game show, contestants choose a number from 1 to 100. If a contestant has chosen number 45, what is the probability that this number is drawn if the game draws 5 numbers randomly without replacement?

6.7 Concluding Remarks

In this chapter, we have examined key discrete probability distributions that are integral to data science: the Bernoulli, Binomial, and Poisson distributions. Understanding these distributions equips data scientists with powerful tools for analyzing categorical data and modeling discrete events.

Mastering these discrete distributions enhances our ability to model and interpret data effectively, leading to more accurate predictions and insights. As

DRAFT

we transition to continuous probability distributions in the next chapter, we will expand our analytical toolkit to handle a broader range of data types and modeling scenarios.

6.8 Chapter Exercises

1. A coin is flipped once. Define the random variable X as the outcome of the coin flip (1 for heads, 0 for tails). Calculate the mean and variance of X.

2. Suppose a factory produces light bulbs, and each bulb has a 90% probability of being functional. If you randomly select one bulb, what is the probability that it is functional?

3. A basketball player has a free throw success rate of 75%. If she takes 10 free throws, what is the probability that she makes exactly 8 of them? (Use the binomial probability formula.)

4. In a survey, it is found that 60% of people prefer coffee over tea. If you randomly sample 15 people, what is the probability that exactly 9 prefer coffee?

5. A call center receives an average of 4 calls per hour. What is the probability that they receive exactly 6 calls in the next hour?

6. In a certain city, the average number of accidents at a particular intersection is 2 per month. What is the probability that there will be no accidents in the next month?

7. A six-sided die is rolled. Define the random variable Y as the outcome of the roll. Calculate the mean and variance of Y.

8. A spinner is divided into 8 equal sections numbered from 1 to 8. What is the probability of landing on an even number when the spinner is spun once?

9. In a quality control process, the probability of finding a defective product is 0.02. If a batch contains 100 products, calculate the probability of finding exactly 3 defective products.

10. An average of 3 earthquakes occur in a year in a certain region. Using the Poisson distribution, calculate the probability of experiencing at least one earthquake in the next year.

11. A game involves rolling a die and flipping a coin. If the die shows a 3 or a 4, you win a prize with a probability of 0.5 (coin shows heads). If the die shows any other number, you don't win. What is the overall probability of winning a prize?

DRAFT

12. A biased coin lands on heads with a probability of 0.65.

    (a) What is the expected number of flips until the coin lands on heads for the first time?

    (b) What is the probability that the first time the coin lands on heads is on the sixth flip?

    (c) If the coin is flipped 5 times,

        (i) What is the probability of getting exactly 4 heads?

        (ii) What is the probability of getting at most 1 head?

13. A quality control inspector checks items on a production line. Each item has a 15% chance of being defective and is discarded if defective.

    (a) What is the expected number of items inspected before a non-defective one is found?

    (b) What is the probability that the fourth item inspected is the first defective one?

    (c) What is the probability that exactly 5 defective items are found among 10 inspected?

14. A call center agent receives calls where each call has a 10% chance of being a complaint.

    (a) What is the probability that the first complaint is received on the 6^{th} call?

    (b) What is the expected number of calls until a complaint is received?

    (c) What is the probability that 2 complaints are received within the first 5 calls?

DRAFT

Chapter 7

Some Continuous Probability Distributions

DRAFT

7.1 Introduction

In probability theory and statistics, continuous probability distributions play a fundamental role in modeling and analyzing real-world phenomena. Unlike discrete distributions, which are defined for countable outcomes, continuous distributions are used to describe outcomes that can take on any value within a given range. This chapter delves into some of the most widely used continuous probability distributions, including the Uniform, Exponential, and Normal distributions.

Continuous probability distributions are integral to various fields, such as engineering, economics, and the natural sciences, due to their ability to represent diverse processes and events accurately. Understanding these distributions enables us to calculate probabilities and make inferences about populations based on sample data.

We begin with the Uniform distribution, which serves as a simple model for random variables that have equally likely outcomes over a specific interval. Following this, we explore the Exponential distribution, commonly used to model the time between events in a Poisson process. We then delve into the Normal distribution, arguably the most important distribution in statistics, due to the Central Limit Theorem's implication that it approximates many natural phenomena.

Each section will provide a detailed definition of the distribution, its properties, and practical examples to illustrate its application. Additionally, exercises are included to reinforce the concepts and allow for hands-on practice in calcu-

lating probabilities and understanding the distribution's behavior.

7.2 Continuous Uniform Distribution

The continuous uniform distribution is a probability distribution that describes an outcome where all values within a specified range are equally likely to occur. This distribution is defined over an interval $[a, b]$, where a is the minimum value and b is the maximum value. In contrast to discrete distributions, which deal with distinct outcomes, the continuous uniform distribution addresses scenarios where the outcomes can take on any value within the range.

The continuous uniform distribution is commonly used in simulations, random sampling, and scenarios where a uniform distribution of outcomes is assumed, such as in generating random numbers or modeling processes where each outcome within a specified range is equally probable. Its simplicity and intuitive nature make it a fundamental concept in statistics and probability theory.

Definition: A random variable X is said to follow a continuous uniform distribution on the interval $[a, b]$, denoted $X \sim U(a, b)$, if its probability density function (PDF) is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The plot of $X \sim U(a, b)$ where X is uniformly distributed between a and b are presented in Figure 7.1.

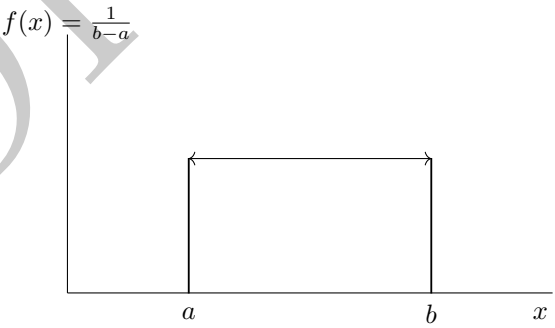


Figure 7.1: The plot of $X \sim U(a, b)$.

DRAFT

7.2.1 Distributional Properties

Expected Value (Mean)

The expected value $E(X)$ is given by

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_a^b x \cdot \frac{1}{b-a}\,dx. \\ &= \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}. \end{aligned}$$

Hence,

$$\text{Mean} = E(X) = \frac{a+b}{2}$$

Variance and Standard Deviation

Expected Value: We already have the expected value $E(X)$ for a uniform random variable X over $[a, b]$

$$E(X) = \frac{a+b}{2}.$$

Expected Value of X^2: To find the variance, we first need $E(X^2)$. This is given by

$$\begin{aligned} E(X^2) &= \int_a^b x^2 f_X(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx \\ &= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b \\ &= \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{b^2+ab+a^2}{3}. \end{aligned}$$

Hence, the variance is

$$\begin{aligned} \text{Var}(X) &= \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

DRAFT

and the standard deviation is

$$\text{Standard Deviation} = \frac{b-a}{\sqrt{12}}.$$

Problem 7.1. *Suppose the time (in minutes) it takes for a customer to be served at a coffee shop follows a continuous uniform distribution between 5 and 15 minutes.*

(a). What is the probability that a customer is served within 10 minutes?

(b). What is the expected time for a customer to be served?

(c). What is the variance of the time for a customer to be served?

Solution

(a). **Probability that a customer is served within 10 minutes:**

Let X be the time taken to be served, and $X \sim \text{Uniform}(5, 15)$. To find $P(X \leq 10)$, the PDF of X is given by

$$f(x) = \frac{1}{15-5} = \frac{1}{10} \text{ for } 5 \leq x \leq 15$$

The probability is given by the integral of the PDF from 5 to 10

$$P(X \leq 10) = \int_5^{10} \frac{1}{10}\, dx = \frac{10-5}{10} = 0.5$$

(b). **Expected time:**

The expected value for a continuous uniform distribution $X \sim \text{Uniform}(a, b)$ is

$$E(X) = \frac{a+b}{2}$$

Here, $a = 5$ and $b = 15$

$$E(X) = \frac{5+15}{2} = 10 \text{ minutes}$$

(c). **Variance of the time:**

The variance for a continuous uniform distribution $X \sim \text{Uniform}(a, b)$ is

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Here, $a = 5$ and $b = 15$, the variance is

$$\text{Var}(X) = \frac{(15-5)^2}{12} = \frac{100}{12} \approx 8.33$$

DRAFT

□

Problem 7.2. *Suppose you are designing a random number generator that outputs a number between 1 and 100. Each number in this range is equally likely to be selected.*

(a). What is the probability that the generator outputs a number between 20 and 50?

(b). Determine the mean and variance of the numbers generated by this random number generator.

(c). If you generate 10,000 numbers, what is the expected number of times a number between 20 and 50 is generated?

Solution

(a). Probability Calculation

Since the numbers generated are uniformly distributed between 1 and 100, we can model this using a continuous uniform distribution $U(a, b)$ with $a = 1$ and $b = 100$.

The probability density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $f(x) = \frac{1}{100-1} = \frac{1}{99}$ for $1 \leq x \leq 100$.

To find the probability that the generator outputs a number between 20 and 50, we calculate

$$\begin{aligned} P(20 \leq X \leq 50) &= \int_{20}^{50} f(x)\,dx = \int_{20}^{50} \frac{1}{99}\,dx \\ &= \frac{1}{99}\int_{20}^{50} dx \\ &= \frac{50-20}{99} \\ &\approx 0.303. \end{aligned}$$

(b). Mean and Variance

For a uniform distribution $U(a, b)$,

$$\text{Mean} = \mu = \frac{a+b}{2} = \frac{1+100}{2} = 50.5,$$

$$\text{Variance} = \sigma^2 = \frac{(b-a)^2}{12} = \frac{(100-1)^2}{12} = \frac{99^2}{12} = \frac{9801}{12} \approx 816.75.$$

DRAFT

(c). Expected Number of Times a Number Between 20 and 50 is Generated

The expected number of times a number between 20 and 50 is generated out of 10,000 numbers can be found by multiplying the probability by the total number of trials

$$E(\text{number of times}) = P(20 \leq X \leq 50) \times 10000 = 0.303 \times 10000 = 3030.$$

Thus, we expect the number generator to output a number between 20 and 50 approximately 3030 times out of 10,000 trials.

7.2.2 Python Code for Uniform Distribution Characteristics

The Uniform distribution models scenarios where all outcomes are equally likely within a given interval. Below is Python code demonstrating how to compute various characteristics for both Continuous and Discrete Uniform distributions.

Continuous Uniform Distribution

```
import numpy as np
from scipy.stats import uniform

# Define the parameters
a = 0  # Lower bound
b = 10  # Upper bound

# Define the Continuous Uniform distribution
dist_continuous = uniform(loc=a, scale=b-a)

# 1. Probability Density Function (PDF)
x_values = np.linspace(a, b, 100)
pdf_values = dist_continuous.pdf(x_values)
print("PDF values for x from {} to {}:".format(a, b),
pdf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist_continuous.cdf(x_values)
print("CDF values for x from {} to {}:".format(a, b),
cdf_values)

# 3. Mean (Expected Value)
mean_continuous = dist_continuous.mean()
print("Mean (Expected Value):", mean_continuous)

# 4. Variance
variance_continuous = dist_continuous.var()
```

```
print("Variance:", variance_continuous)

# 5. Standard Deviation
std_dev_continuous = dist_continuous.std()
print("Standard Deviation:", std_dev_continuous)

# 6. Quantiles
quantiles_continuous = dist_continuous.ppf([0.25, 0.5, 0.75])  # 25th, 50th (median), and 75th percentiles
print("Quantiles at 0.25, 0.5, and 0.75:", quantiles_continuous)

# 7. Percentiles
percentiles_continuous = dist_continuous.ppf([0.1, 0.9])  # 10th and 90th percentiles
print("Percentiles at 0.1 and 0.9:", percentiles_continuous)

```

7.2.3 Exercises

1. Consider a discrete uniform distribution where X takes values $\{1, 2, 3, \ldots, n\}$. (i) Show that the expected value $E(X)$ is $\frac{n+1}{2}$ and (ii) Find the Variance of X.

2. Find the cumulative distribution function (CDF) of a continuous uniform random variable $X \sim U(a, b)$.

3. If $X \sim U(0, 1)$, find the distribution of $Y = a + (b - a)X$.

4. Suppose the waiting time for a bus is uniformly distributed between 0 and 30 minutes. What is the probability that a person will wait more than 20 minutes?

5. A factory produces items with weights that are uniformly distributed between 50 grams and 150 grams. What is the probability that a randomly chosen item weighs between 80 grams and 120 grams?

6. The cholesterol level of adults in a certain region follows a uniform distribution between 150 mg/dL and 250 mg/dL.

   (a) Write the probability density function $f(x)$ of the cholesterol level.

   (b) What is the probability that a randomly selected adult has a cholesterol level between 180 mg/dL and 220 mg/dL?

   (c) Find the mean and variance of the cholesterol level.

7. A machine produces metal rods that are uniformly distributed in length between 98 cm and 102 cm. What is the probability that a randomly selected rod is between 99 cm and 101 cm in length?

DRAFT

8. In a quality control process, the time to inspect a product is uniformly distributed between 1 and 5 minutes. Find the probability that the inspection time for a randomly chosen product is more than 4 minutes.

9. The download speed of a certain internet connection is uniformly distributed between 10 Mbps and 100 Mbps. What is the probability that the download speed at any given time is less than 50 Mbps?

10. The delivery time for a package from a warehouse to a customer is uniformly distributed between 2 and 7 days. What is the probability that a package will be delivered in less than 4 days?

11. The time a wildlife photographer waits to see a particular bird is uniformly distributed between 30 minutes and 3 hours. Find the probability that the wait time is more than 2 hours.

12. The fuel efficiency of a car is uniformly distributed between 15 and 25 miles per gallon. What is the probability that the car's fuel efficiency is between 18 and 22 miles per gallon?

7.3 Exponential Distribution

Imagine you are at a busy city street waiting for a taxi. Taxis arrive at the street randomly but at an average rate of 5 taxis per hour. You're curious about how long you might have to wait for the next taxi. Is there a way to predict or understand the waiting time better?

You notice that sometimes the wait is short, and other times it can be quite long. This variability and randomness in waiting times suggest a need for a mathematical model to describe it.

The **exponential distribution** is a probability distribution used to model the time between successive events in a process where events occur continuously and independently at a constant average rate. It is characterized by its **rate parameter** λ, which is the reciprocal of the average time between events. It is widely applied in various fields, including data science, to analyze the duration or time between events. In our taxi example, it helps us understand and predict the waiting time until the next taxi arrives. The exponential distribution provides a framework to calculate probabilities and make informed decisions based on the average rate of arrivals.

Definition: The continuous random variable, say X is said to have an

DRAFT

exponential distribution, if it has the following probability density function

$$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0, \end{cases} \tag{7.1}$$

where $\lambda > 0$ is the rate parameter.

The density plot of X is present in Figure 7.2 for value of $\lambda = 0.04$.

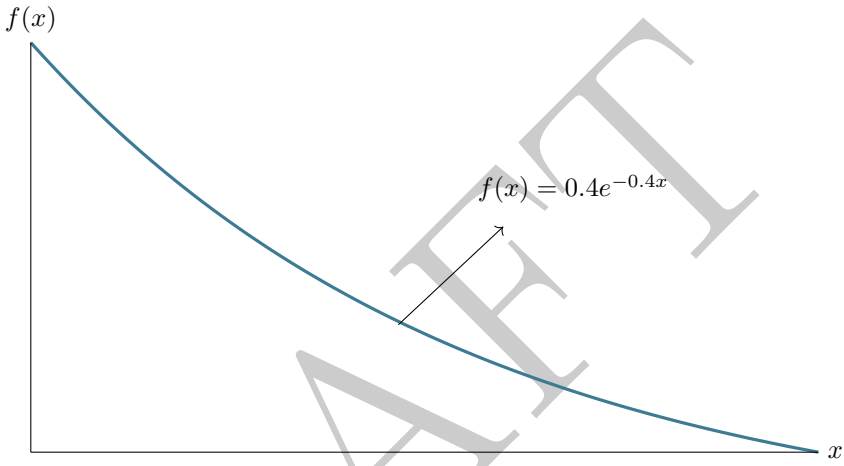


Figure 7.2: The probability density function $f(x) = 0.4e^{0.4x}$.

This probability is the area under the probability density function between the points $a = 1$ and $b = 2$ as illustrated in Figure 7.3.

DRAFT

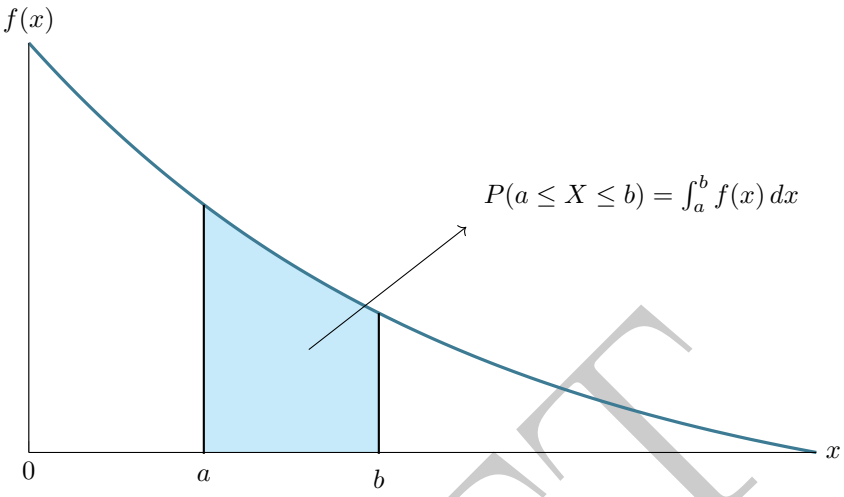


Figure 7.3: The area under the probability density function $f(x)$ between a and b.

The CDF is obtained by integrating the PDF

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t}\,dt \\ &= -e^{-\lambda t}\Big|_0^x \\ &= -e^{-\lambda x} + e^0 \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Thus, the CDF $F(x)$ for the exponential distribution is:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

The CDF of the exponential distribution for $\lambda = 0.4$ is $F(x) = 1 - e^{-0.4x}$ and the graphical presentation is presented in Figure 7.4.

DRAFT

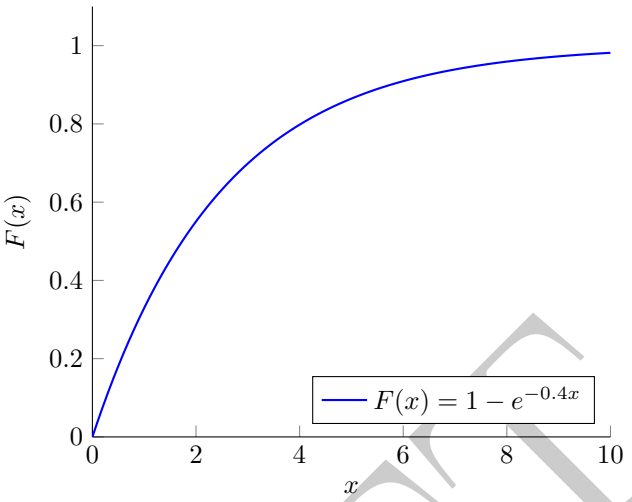


Figure 7.4: CDF of the Exponential Distribution with $\lambda = 0.4$

Theorem 7.1. *Let X be a random variable that follows an exponential distribution with parameter λ. The rth mean, defined as $E[X^r]$, is given by*

$$E[X^r] = \frac{r!}{\lambda^r}$$

Proof. The probability density function (PDF) of the exponential distribution is given by:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

To find $E[X^r]$, we compute

$$E[X^r] = \int_0^\infty x^r f(x)\, dx = \int_0^\infty x^r \lambda e^{-\lambda x}\, dx.$$

Using the substitution $u = \lambda x$ (hence $x = \frac{u}{\lambda}$ and $dx = \frac{du}{\lambda}$), we have

$$E[X^r] = \int_0^\infty \left(\frac{u}{\lambda}\right)^r \lambda e^{-u} \frac{du}{\lambda} = \frac{\lambda}{\lambda^{r+1}} \int_0^\infty u^r e^{-u}\, du.$$

The integral $\int_0^\infty u^r e^{-u}\, du$ is the definition of the gamma function $\Gamma(r+1)$, which equals $r!$

$$E[X^r] = \frac{\lambda}{\lambda^{r+1}} \cdot r! = \frac{r!}{\lambda^r}.$$

This completes the proof for the rth mean of the exponential distribution. □

DRAFT

Mean and Variance of the Exponential Distribution

Using Theorem 7.1, we can easily find the mean and variance of the exponential distribution.

Mean

For the exponential distribution (specifically when $r = 1$)

$$E[X] = E[X^1] = \frac{1!}{\lambda^1} = \frac{1}{\lambda}$$

Variance

The variance $\text{Var}(X)$ is calculated as follows

$$\text{Var}(X) = E[X^2] - (E[X])^2.$$

Calculating $E[X^2]$ (using $r = 2$), we have

$$E[X^2] = \frac{2!}{\lambda^2} = \frac{2}{\lambda^2}$$

Now, substituting into the variance formula

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Problem 7.3. *A data center experiences server failures where the time between failures follows an exponential distribution with a mean of 10 days.*

(i). Calculate the rate parameter λ.

(ii). What is the probability that a server will fail within the next 3 days?

(iii). What is the probability that the server will last longer than 12 days without failing?

(i). Calculate the Rate Parameter λ

The mean time between failures is given as 10 days. The rate parameter λ is calculated using the formula

$$\lambda = \frac{1}{\text{Mean}} = \frac{1}{10} = 0.1.$$

(ii). Probability of Failure Within the Next 3 Days

To find the probability that a server will fail within the next 3 days, we use the cumulative distribution function (CDF) of the exponential distribution

$$P(X \leq 3) = 1 - e^{-\lambda \times 3} = 1 - e^{-0.1 \times 3} = 1 - e^{-0.3} = 1 - 0.7408 = 0.2592.$$

So, the probability that a server will fail within the next 3 days is approximately 0.2592 (or 25.92%).

DRAFT

(iii). Probability of Lasting Longer Than 12 Days

To find the probability that the server will last longer than 12 days, we can use the survival function

$$P(X > 12) = 1 - P(X \leq 12) = e^{-\lambda \times 12} = e^{-0.1 \times 12} = e^{-1.2} \approx 0.3012.$$

So, the probability that the server will last longer than 12 days without failing is approximately 0.3012 (or 30.12%).

Problem 7.4. *In a hospital, the time between arrivals of patients in the emergency room follows an exponential distribution with an average time of 15 minutes.*

(a) What is the probability that the time between two successive arrivals is more than 20 minutes?

(b) What is the probability that the time between two successive arrivals is less than 10 minutes?

(c) Calculate the expected time between two successive arrivals and its standard deviation.

Solution

(a). Let X be the time between arrivals, which follows an exponential distribution with parameter λ. The rate parameter λ is the reciprocal of the mean, so $\lambda = \frac{1}{15}$ per minute. The probability that the time between two successive arrivals is more than 20 minutes is calculated as follows

$$P(X > 20) = 1 - P(X \leq 20) = 1 - F(20) = 1 - \left(1 - e^{-\lambda \cdot 20}\right) = e^{-\lambda \cdot 20}$$

Substituting $\lambda = \frac{1}{15}$,

$$P(X > 20) = e^{-\frac{20}{15}} = e^{-\frac{4}{3}} \approx 0.2636$$

Thus, the probability that the time between two successive arrivals is more than 20 minutes is approximately 0.2636.

(b). The probability that the time between two successive arrivals is less than 10 minutes is calculated as follows

$$P(X < 10) = F(10) = 1 - e^{-\lambda \cdot 10}$$

Substituting $\lambda = \frac{1}{15}$,

$$P(X < 10) = 1 - e^{-\frac{10}{15}} = 1 - e^{-\frac{2}{3}} \approx 0.4866$$

Thus, the probability that the time between two successive arrivals is less than 10 minutes is approximately 0.4866.

DRAFT

(c). The expected time between two successive arrivals (the mean of the exponential distribution) is given by

$$E(X) = \frac{1}{\lambda} = 15 \text{ minutes}$$

The standard deviation of the time between two successive arrivals is the same as the mean for an exponential distribution, so

$$\text{Standard deviation} = \frac{1}{\lambda} = 15 \text{ minutes}$$

Therefore, the expected time between two successive arrivals is 15 minutes, and the standard deviation is also 15 minutes.

□

Theorem 7.2. *The characteristic function of an exponential random variable X with rate parameter λ is given by*

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}, \quad \textit{for } t \in \mathbb{R}.$$

Proof. To find the characteristic function $\varphi_X(t)$, we need to calculate the expected value $E[e^{itX}]$

$$\varphi_X(t) = E[e^{itX}] = \int_0^\infty e^{itx} \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{-(\lambda - it)x}\, dx.$$

The integral is of the form $\int_0^\infty e^{-ax}\, dx = \frac{1}{a}$ for $\Re(a) > 0$

$$\varphi_X(t) = \lambda \left[\frac{1}{\lambda - it} \right] = \frac{\lambda}{\lambda - it}.$$

Thus, the characteristic function of X is

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}.$$

□

Theorem 7.3. *The moment generating function (MGF) of an exponential random variable X with rate parameter λ is given by*

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad \textit{for } t < \lambda.$$

DRAFT

Proof. To find the moment generating function $M_X(t)$, we need to calculate the expected value $E[e^{tX}]$

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\,dx = \lambda\left[\frac{1}{\lambda - t}\right].$$

Thus, the moment generating function of X is

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad \text{for } t < \lambda.$$

□

7.3.1 Memoryless Property

In probability and statistics, the concept of memorylessness refers to a characteristic of specific probability distributions. A random variable X is said to have the memoryless property if the probability of an event occurring in the future is independent of any past events. In other words, it "forgets" what has happened before. Imagine you're waiting for a bus that comes randomly every few minutes. If you've been waiting for 10 minutes and the bus hasn't arrived yet, the probability of it arriving in the next 5 minutes is the same as if you had just started waiting.

Only two types of distributions exhibit the memoryless property: geometric and exponential probability distributions. The exponential distribution has the memoryless property, which states that the probability of an event occurring in the next t units of time is independent of how much time has already elapsed. Mathematically, this can be expressed in the following theorem.

Theorem 7.4 (Memoryless Property)**.** *Let X be an exponentially distributed random variable with rate parameter $\lambda > 0$. Then, for all $s, t \geq 0$,*

$$P(X > s + t \mid X > s) = P(X > t).$$

Proof. The proof of the memoryless property relies on the definition of conditional probability and the exponential distribution's probability density function.

By definition of conditional probability, we have

$$P(X > s + t \mid X > s) = \frac{P(X > s + t \text{ and } X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)}.$$

Since X is exponentially distributed with rate parameter λ, the cumulative distribution function (CDF) is

$$F(x) = 1 - e^{-\lambda x} \quad \text{for } x \geq 0,$$

DRAFT

and the survival function (which gives the probability that X is greater than a certain value) is

$$P(X > x) = 1 - F(x) = e^{-\lambda x}.$$

Using the survival function, we can rewrite the conditional probability

$$P(X > s + t \mid X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).$$

Thus, the memoryless property is proved. □

Problem 7.5. *Consider a system with a component whose time to failure is exponentially distributed with a rate parameter $\lambda = 0.01$ failures per hour. If the component has been operational for 100 hours without failure, what is the probability that it will operate for at least another 50 hours?*

Solution

To solve this problem, we use the properties of the exponential distribution, specifically its memoryless property. Let X represent the time to failure of the component. For an exponentially distributed random variable $X \sim \text{Exp}(\lambda)$, we use the memoryless property which is

$$P(X > s + t \mid X > s) = P(X > t).$$

Here, we need to find the probability that the component will operate for at least 50 more hours, given that it has already operated for 100 hours. Using the memoryless property

$$s = 100, \qquad t = 50, \qquad \lambda = 0.01.$$

Thus,

$$P(X > 100 + 50 \mid X > 100) = P(X > 50) = e^{-\lambda \times 50} = e^{-0.01 \times 50} \approx 0.6065.$$

Therefore, the probability that the component will operate for at least another 50 hours is approximately 0.6065.

□

Problem 7.6. *Suppose that the waiting time for a bus at a certain bus stop is exponentially distributed with a mean waiting time of 10 minutes. Given that a person has already waited for 5 minutes, what is the probability that they will have to wait at least an additional 10 minutes?*

DRAFT

Solution

. Let X denote the waiting time. The mean waiting time is 10 minutes, so the rate parameter is

$$\lambda = \frac{1}{10}.$$

Using the memoryless property of the exponential distribution:

$$P(X > s + t \mid X > s) = P(X > t).$$

Here,

$$s = 5, \qquad t = 10.$$

Thus,

$$P(X > 5 + 10 \mid X > 5) = P(X > 10) = e^{-\lambda \times 10} = e^{-(\frac{1}{10})(10)} = e^{-1} \approx 0.3679.$$

Therefore, the probability that the person will have to wait at least an additional 10 minutes is 0.3679. □

7.3.2 Properties of the Exponential Distribution

Let X be an exponential random variable with rate parameter λ. The following are the key properties of the exponential distribution:

- **Cumulative Distribution Function (CDF):**

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

- **Mean (Expected Value):**

$$E[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}.$$

- **Variance:**

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2}.$$

- **Standard Deviation:**

$$\sigma_X = \frac{1}{\lambda}.$$

- **Memoryless Property:** The exponential distribution has the memoryless property, which states that

$$P(X > s + t \mid X > s) = P(X > t) \quad \text{for all } s, t \geq 0.$$

DRAFT

- **Moment Generating Function (MGF):**

$$M_X(t) = E[e^{tX}] = \frac{\lambda}{\lambda - t}, \quad \text{for } t < \lambda.$$

- **Characteristic Function:**

$$\varphi_X(t) = E[e^{itX}] = \frac{\lambda}{\lambda - it}, \quad \text{for } t \in \mathbb{R}.$$

- **Quantile Function:** The quantile function (inverse of the CDF) for $0 < p < 1$ is given by:

$$Q(p) = F^{-1}(p) = -\frac{1}{\lambda}\ln(1-p).$$

- **Relationship with the Poisson Process:** If $X \sim \text{Exponential}(\lambda)$, it can be interpreted as the waiting time between events in a Poisson process with rate λ.

- **Sum of Independent Exponential Variables:** The sum of n independent exponential random variables with the same rate parameter λ follows a Gamma distribution

$$Y = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

  For $n = 1$, the Gamma distribution reduces to the exponential distribution.

- **Relationship with Other Distributions:**
  - If $X \sim \text{Exponential}(\lambda)$, then $X \sim \text{Gamma}(1, \lambda)$.
  - If $X \sim \text{Exponential}(\lambda)$, then $Y = \frac{X}{\beta} \sim \text{Exponential}\left(\frac{\lambda}{\beta}\right)$ for $\beta > 0$.

Applications in Data Science

The important applications of exponential distribution are mentioned in the following:

1. **Survival Analysis**: In medical research and clinical trials, the exponential distribution is used to model the time until an event occurs, such as patient survival times or time to recovery from a treatment.

2. **Queuing Theory**: Exponential distributions model the time between arrivals in queuing systems, like customer service lines, call centers, and computer networks. This helps in analyzing and optimizing service processes and reducing wait times

DRAFT

3. **Reliability Engineering**: It helps in estimating the lifespan of products and systems, providing insights into maintenance schedules and warranty analysis.

4. **Markov Processes**: As part of continuous-time Markov chains, exponential distributions model the time spent in each state before transitioning to another state.

5. **Customer Behavior Analysis**: Exponential models can help analyze customer behavior, such as the time between purchases or interactions with a service, allowing businesses to tailor marketing strategies effectively.

7.3.3 Python Code for Exponential Distribution Characteristics

The Exponential distribution models the time between events in a Poisson process. Below is Python code demonstrating how to compute various characteristics of the Exponential distribution.

Python Code

```
import numpy as np
from scipy.stats import expon

# Define the parameter lambda (rate of occurrence)
lambda_ = 1  # Rate parameter
scale = 1 / lambda_  # Scale parameter for scipy's expon

# Exponential distribution
dist = expon(scale=scale)

# 1. Probability Density Function (PDF)
x_values = np.linspace(0, 10, 100)  # Values for the
random variable
pdf_values = dist.pdf(x_values)
print("PDF values for x from 0 to 10:", pdf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist.cdf(x_values)
print("CDF values for x from 0 to 10:", cdf_values)

# 3. Mean (Expected Value)
mean = dist.mean()
print("Mean (Expected Value):", mean)

# 4. Variance
```

```
    variance = dist.var()
    print("Variance:", variance)

    # 5. Standard Deviation
    std_dev = dist.std()
    print("Standard Deviation:", std_dev)

    # 6. Quantiles
    quantiles = dist.ppf([0.25, 0.5, 0.75])  # 25th, 50th (median), and 75th percentiles
    print("Quantiles at 0.25, 0.5, and 0.75:", quantiles)

    # 7. Percentiles
    percentiles = dist.ppf([0.1, 0.9])  # 10th and 90th percentiles
    print("Percentiles at 0.1 and 0.9:", percentiles)

    # 8. Moment Generating Function (MGF)
    def mgf(t, lambda_):
    return 1 / (1 - t / lambda_)  # MGF of Exponential distribution for t < lambda_

    # MGF values at t = 0.1 and t = 0.5
    mgf_values = [mgf(t, lambda_) for t in [0.1, 0.5]]
    print("MGF values for t = 0.1 and t = 0.5:", mgf_values)

    # 9. Probability Generating Function (PGF)
    # Note: The PGF is not typically used for continuous distributions like the Exponential distribution.
    # Here, we will use the MGF as a substitute.
```

Explanations

- **Probability Density Function (PDF):**
  - The PDF provides the likelihood of each value. Compute these values using `dist.pdf(x_values)`.
- **Cumulative Distribution Function (CDF):**
  - The CDF gives the probability that the random variable is less than or equal to a given value. Compute these values using `dist.cdf(x_values)`.
- **Mean (Expected Value):**
  - The mean (expected value) is $\frac{1}{\lambda}$. Compute this using `dist.mean()`.

- **Variance:**
  - The variance is $\frac{1}{\lambda^2}$. Compute this using `dist.var()`.
- **Standard Deviation:**
  - The standard deviation is $\frac{1}{\lambda}$. Compute this using `dist.std()`.
- **Quantiles:**
  - Quantiles are values below which a given proportion of data falls. Compute these using `dist.ppf()` for the desired quantile probabilities.
- **Percentiles:**
  - Percentiles are specific quantiles, such as the 10th and 90th percentiles. Compute these using `dist.ppf()` for the desired percentiles.
- **Moment Generating Function (MGF):**
  - The MGF for the Exponential distribution is $M_X(t) = \frac{1}{1-t/\lambda}$ for $t < \lambda$. Define a function `mgf(t, lambda_)` to compute MGF values.
- **Probability Generating Function (PGF):**
  - The PGF is generally used for discrete distributions. For continuous distributions like the Exponential, the MGF serves a similar purpose.

7.3.4 Exercises

1. The lifetime of a particular brand of lightbulb is exponentially distributed with a mean of 1000 hours.
   (a) What is the probability that a lightbulb lasts more than 1200 hours?
   (b) What is the probability that a lightbulb lasts between 800 and 1200 hours?
2. A machine in a factory breaks down on average once every 500 hours. The time between breakdowns is exponentially distributed.
   (a) What is the probability that the machine will operate for at least 1000 hours without a breakdown?
   (b) Determine the probability that the machine will break down within the next 200 hours.

DRAFT

3. The waiting time for a specific genetic test result from a laboratory follows an exponential distribution with a mean of 2 days.

   (a) Find the probability that the test result will be available in less than 1 day.

   (b) Find the probability that the test result will take more than 3 days.

4. The time until failure of a critical component in a medical device follows an exponential distribution with a mean of 5 years.

   (a) Calculate the probability that the component will fail within the first 3 years.

   (b) Determine the probability that the component will last more than 7 years.

5. In a call center, the time between consecutive calls follows an exponential distribution with a mean time of 4 minutes.

   (a) What is the probability that the next call comes within 2 minutes?

   (b) What is the probability that the next call will not come for at least 6 minutes?

6. The time between arrivals of buses at a particular bus stop follows an exponential distribution with an average time of 20 minutes.

   (a) What is the probability that a bus will arrive in the next 5 minutes?

   (b) Calculate the probability that you will have to wait more than 25 minutes for the next bus.

7. The lifespan of a certain species of bacteria follows an exponential distribution with a mean of 24 hours.

   (a) Find the probability that a bacterium lives longer than 30 hours.

   (b) Determine the probability that a bacterium lives between 10 and 20 hours.

8. The response time of a server to a network request is exponentially distributed with an average time of 0.5 seconds.

   (a) What is the probability that the server responds in less than 0.3 seconds?

   (b) Calculate the probability that the server takes more than 1 second to respond.

9. The time it takes for a chemical reaction to complete in a lab experiment follows an exponential distribution with a mean of 45 minutes.

DRAFT

(a) What is the probability that the reaction will complete in less than 30 minutes?

(b) What is the probability that the reaction will take more than 60 minutes to complete?

10. The lifespan of a certain species of laboratory mice follows an exponential distribution with a mean lifespan of 2.5 years.

(a) Determine the probability density function of the lifespan.

(b) What is the median lifespan of the mice?

(c) Calculate the variance and standard deviation of the lifespan.

(d) What is the probability that a randomly selected mouse lives more than 3 years?

11. The time (in hours) until recovery from a certain disease follows an exponential distribution with a mean of 2 hours.

(a) What is the probability that a patient will recover in less than 1 hour?

(b) What is the probability that a patient will recover between 1 and 3 hours?

(c) Find the median recovery time.

12. Assume that the lifetime of a light bulb follows an exponential distribution with a mean lifetime of 1000 hours. Let X be the lifetime of the light bulb. If a light bulb has already been used for 800 hours, what is the probability that it will last at least an additional 500 hours?

7.4 Normal Distribution

Imagine you are working for a clothing company that wants to optimize its garment sizes. To do this effectively, you need to understand the distribution of body measurements, such as heights, within your target customer base. If you were to collect height data from a large number of individuals, what kind of pattern would you expect to see? Intuitively, you would likely find that most people are of average height, with fewer individuals being exceptionally tall or short. This common phenomenon, where measurements cluster around a central value with a symmetrical spread on either side, is a hallmark of the **normal distribution**, also known as the **Gaussian distribution**..

In the field of data science, the normal distribution plays a pivotal role due to its ubiquitous nature and the mathematical properties that simplify analysis. It is characterized by its **symmetric**, **bell-shaped curve** and is instrumental in various statistical methods, including hypothesis testing, regression analysis,

DRAFT

and many machine learning algorithms.

The normal distribution is especially useful because of the **Central Limit Theorem**, which states that the sum of a large number of independent, identically distributed variables tends toward a normal distribution, regardless of the original distribution of the variables. This makes the normal distribution a powerful tool for modeling real-world phenomena and for making inferences about populations based on sample data.

The normal distribution is characterized by its bell-shaped curve, which is symmetric about its mean. The parameters of the normal distribution are the mean (μ) and the standard deviation (σ). The mean represents the central value of the distribution, while the standard deviation measures the spread of the distribution.

7.4.1 Definition of the Normal Distribution

Normal Distribution: A random variable X is said to follow a normal distribution if its probability density function (PDF) is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

where μ is the mean and σ is the standard deviation of the distribution. The graphical presentation of a normal distribution presented in Figure 7.5. The probability density function is a bell-shaped curve that is symmetric about μ. The notation

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

denotes that the random variable X has a normal distribution with mean μ and variance σ^2.

DRAFT

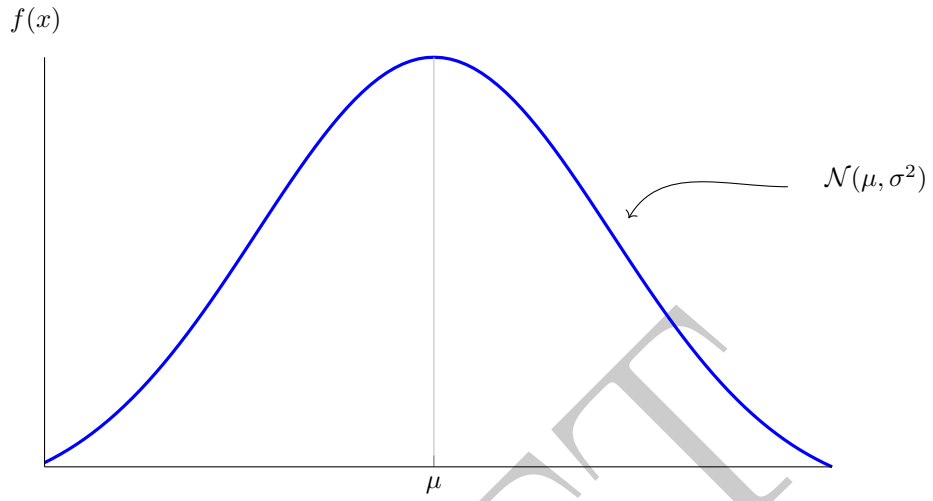


Figure 7.5: The graphical presentation of a normal distribution.

The mean and the variance

$$E(X) = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2$$

of the distribution.

The probability density function of a normal random variable is symmetric around the mean value μ and exhibits a "bell-shaped" curve. Figure 7.6 displays the probability density functions of normal distributions with $\mu = 5, \sigma = 2$ and $\mu = 10, \sigma = 2$. It illustrates that while altering the mean value μ shifts the location of the density function, it does not affect its shape. In contrast, Figure 7.7 presents the probability density functions of normal distributions with $\mu = 5, \sigma = 2$ and $\mu = 5, \sigma = 0.5$. Here, the central position of the density function remains the same, but its shape changes. Larger values of the variance σ^2 lead to wider, flatter bell-shaped curves, whereas smaller values of the variance σ^2 produce narrower, sharper bell-shaped curves.

DRAFT

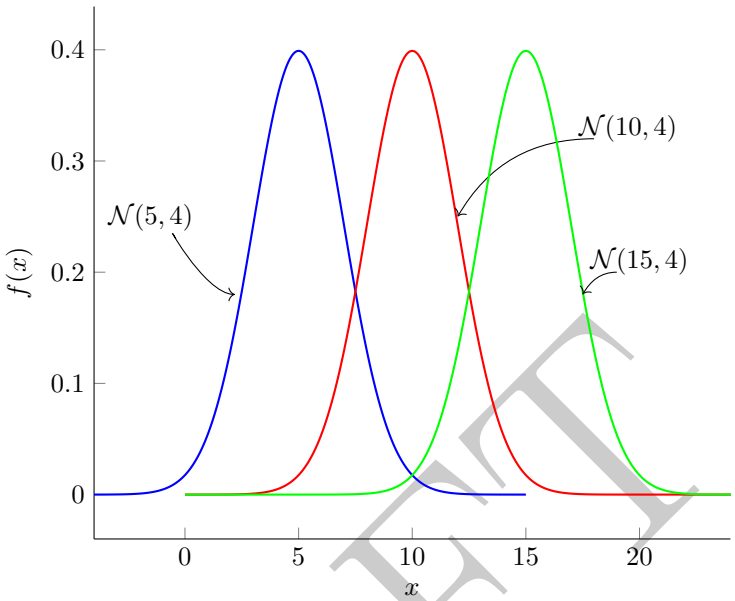


Figure 7.6: The effect of changing the mean of a normal distribution.

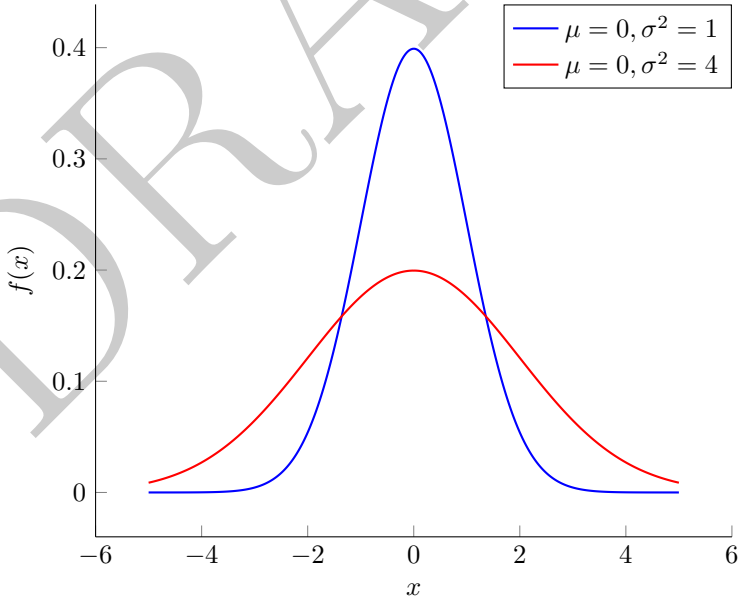


Figure 7.7: The effect of changing the variance of a normal distribution.

DRAFT

7.4.2 Properties of the Normal Distribution

The normal distribution, also known as the **Gaussian distribution**, is one of the most important probability distributions in statistics. The normal distribution has several important properties:

1. **Symmetry:** The normal distribution is symmetric about its mean μ. That is $f(\mu - x) = f(\mu + x)$. This means that the left half of the distribution is a mirror image of the right half.

2. **Bell-Shaped Curve:** The PDF of the normal distribution forms a bell-shaped curve that is unimodal, meaning it has a single peak at the mean μ.

3. **Mean, Median, and Mode:** For a normal distribution, the mean, median, and mode are all equal and located at μ.

4. **Total Area:** The total area under the curve and above the horizontal axis is equal to 1.

5. **Inflection Points:** The points at which the curve changes concavity are located at $\mu - \sigma$ and $\mu + \sigma$.

6. **Empirical Rule (68-95-99.7 Rule):** Approximately 68% of the data lies within one standard deviation of the mean ($\mu \pm \sigma$), about 95% within two standard deviations ($\mu \pm 2\sigma$), and about 99.7% within three standard deviations ($\mu \pm 3\sigma$). See the Figure 7.8.

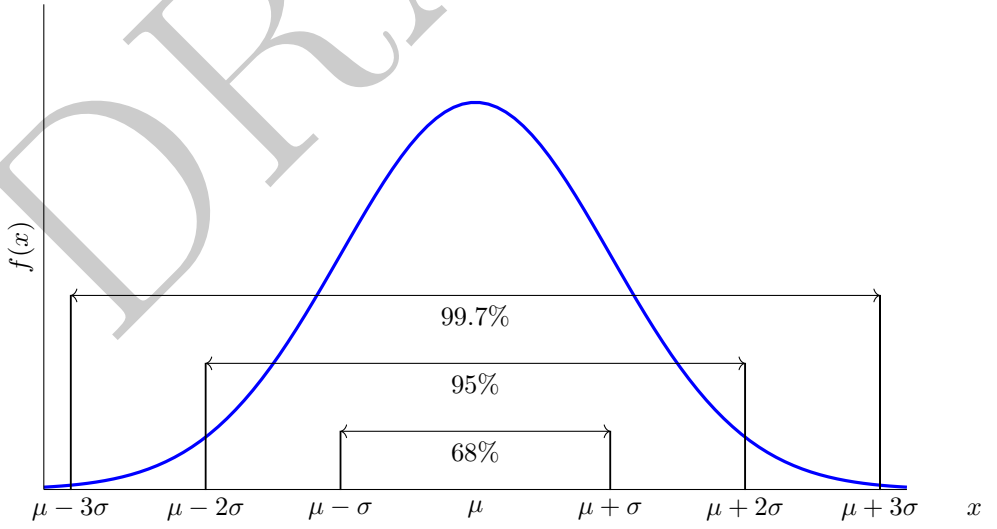


Figure 7.8: The Empirical Rule for Normal Distribution

DRAFT

7. **Asymptotic Behavior:** The tails of the normal distribution approach the horizontal axis but never touch it. This implies that every value, no matter how extreme, has a non-zero probability of occurring.

8. **Linear Transformations:** If X is normally distributed with mean μ and standard deviation σ, then a linear transformation of X given by $Y = aX + b$ is also normally distributed with mean $a\mu + b$ and standard deviation $|a|\sigma$.

9. **Additivity:** The sum of two independent normal random variables is also normally distributed. Specifically, if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

10. **Moment Generating Function:** The moment generating function $M_X(t)$ of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

11. **Characteristic Function:** The characteristic function $\varphi_X(t)$ of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by:

$$\varphi_X(t) = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right).$$

12. **Cumulative Distribution Function (CDF):** The cumulative distribution function (CDF) of the normal distribution with mean μ and standard deviation σ is given by:

$$F(x) = \Pr(X \leq x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right]$$

where $\operatorname{erf}(z)$ is the error function, defined as:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt.$$

Theorem 7.5. *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the mean and standard deviation are μ and σ^2, respectively.*

Proof. To find the expression of mean, we first calculate

$$E(X - \mu) = \int_{-\infty}^{\infty} (x - \mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Setting $z = \frac{x-\mu}{\sigma}$ and $dx = \sigma dz$, we obtain

$$E(X - \mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz = 0,$$

DRAFT

since the integrand above is an odd function of z. Using Theorem 4.5 on page 128, we conclude that

$$E(X) = \mu.$$

The variance of the normal distribution is given by

$$E[(X-\mu)^2] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Again setting $z = \frac{x-\mu}{\sigma}$ and $dx = \sigma dz$, we obtain

$$E[(X-\mu)^2] = \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz.$$

Integrating by parts with $u = z$ and $dv = ze^{-z^2/2}dz$ so that $du = dz$ and $v = -e^{-z^2/2}$, we find that

$$E[(X-\mu)^2] = \sigma^2 \left[\frac{-ze^{-z^2/2}}{\sqrt{2\pi}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2} dz \right] = \sigma^2(0+1) = \sigma^2.$$

□

7.4.3 Inflection Points of the Normal Distribution

Consider the normal probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right).$$

To determine the inflection points, we examine the second derivative of the density. Differentiating twice of $f(x)$ with respect to x, we obtain

$$f''(x) = f(x) \, \frac{(x-\mu)^2 - \sigma^2}{\sigma^4}.$$

The inflection points occur where the second derivative is zero:

$$f''(x) = 0 \quad \Longrightarrow \quad (x-\mu)^2 - \sigma^2 = 0.$$

Solving,

$$(x-\mu)^2 = \sigma^2 \quad \Longrightarrow \quad x = \mu - \sigma \quad \text{and} \quad x = \mu + \sigma.$$

Figure 7.9 shows the second derivative of the standard normal density. The points where the curve crosses zero correspond to the inflection points.

These two points lie exactly one standard deviation below and above the mean. The curvature of the normal curve behaves as follows:

- It is **concave down** on the interval $\mu - \sigma < x < \mu + \sigma$.

DRAFT

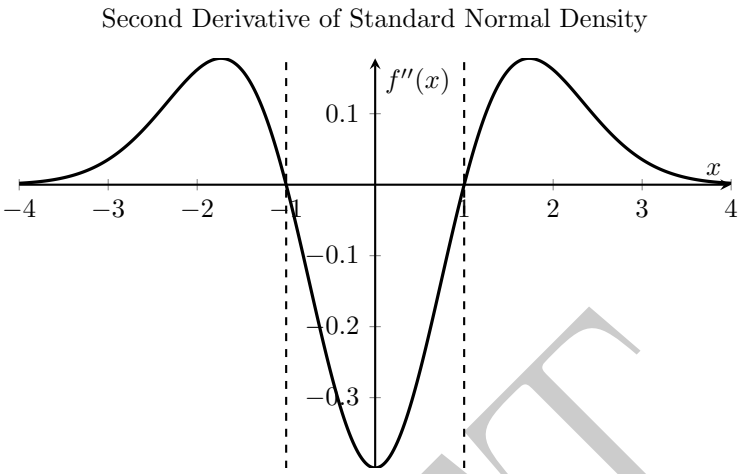


Figure 7.9: Second derivative of the standard normal density.

- It is **concave up** for $x < \mu - \sigma$ and $x > \mu + \sigma$.

Thus, the bell-shaped normal curve changes curvature at the two points

$$x = \mu \pm \sigma.$$

Theorem 7.6. *For a normal distribution with mean μ and standard deviation σ, the inflection points occur at $x = \mu - \sigma$ and $x = \mu + \sigma$.*

Proof. Inflection points occur where the second derivative changes sign, indicating a change in curvature from concave down to concave up. Therefore, to locate the inflection points, we examine the second derivative.

First Derivative
The first derivative of $f(x)$ with respect to x is:

$$f'(x) = \frac{d}{dx}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

Using the chain rule:

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left(-\frac{(x-\mu)}{\sigma^2}\right)$$

Simplifying:

$$f'(x) = -\frac{(x-\mu)}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Second Derivative
The second derivative of $f(x)$ with respect to x is:

$$f''(x) = \frac{d}{dx}\left(-\frac{(x-\mu)}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

DRAFT

Again, using the product rule and chain rule:

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left(1 - \frac{(x-\mu)^2}{\sigma^2}\right)$$

Simplifying:

$$f''(x) = \frac{1}{\sigma^5\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left((x-\mu)^2 - \sigma^2\right)$$

Setting the Second Derivative to Zero
To find the inflection points, we set $f''(x) = 0$:

$$(x-\mu)^2 - \sigma^2 = 0$$

Solving for x,

$$(x-\mu)^2 = \sigma^2$$

or,

$$x - \mu = \pm\sigma$$

$$\therefore x = \mu \pm \sigma$$

The curve bends (changes curvature) at exactly those two points. Thus, the inflection points of the normal distribution are at $x = \mu - \sigma$ and $x = \mu + \sigma$. □

7.4.4 Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. It is denoted by Z and its properties are as follows:

- **Probability Density Function (PDF):** The PDF of the standard normal distribution is given by:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

as illustrated in Figure 7.10.

- **Cumulative Distribution Function:** The CDF of the standard normal distribution is denoted by $\Phi(z)$ and is defined as:

$$\Phi(z) = \Pr(Z \leq z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

DRAFT

where $\operatorname{erf}(z)$ is the error function:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$$

Below is a plot of the standard normal distribution's PDF.

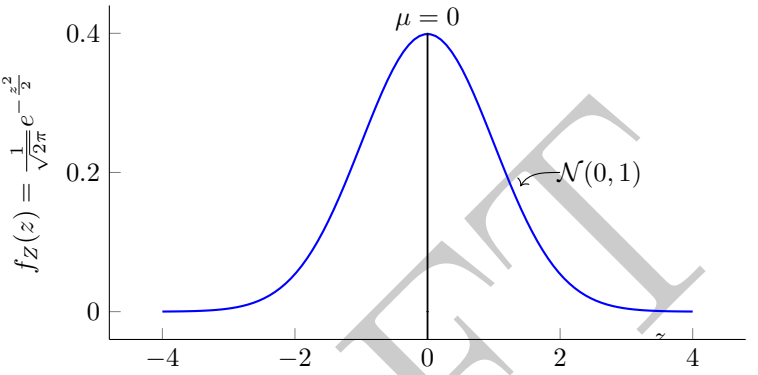


Figure 7.10: The standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

The symmetry of the standard normal distribution about 0 implies that if the random variable Z has a standard normal distribution, then

$$1 - \Phi(z) = \Pr(Z \geq z) = \Pr(Z \leq -z) = \Phi(-z),$$

as illustrated in Figure 7.10. This equation can be rearranged to provide the easily remembered relationship

$$\Phi(z) + \Phi(-z) = 1.$$

The plot presented in Figure 7.11, illustrates the cumulative distribution functions $\Phi(z)$ and $\Phi(-z)$ of the standard normal distribution. The symmetry of the standard normal distribution is evident from these plots.

DRAFT

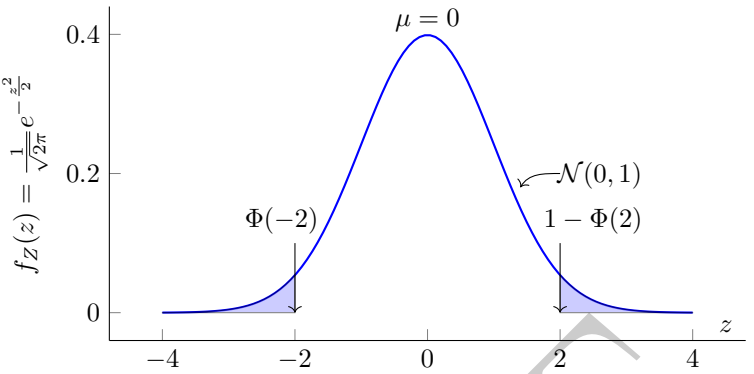


Figure 7.11: Standard normal distribution with shaded areas for $\Phi(-2)$ and $1 - \Phi(2)$.

Problem 7.7. *Suppose* $Z \sim \mathcal{N}(0, 1)$.

(a). Find $P(Z < 1.25)$.

(b). Compute $P(Z > 0.85)$.

(c). Find $P(-1.2 < Z < 0.5)$.

(d). Compute $P(|Z| < 2)$.

(e). Find z *such that* $P(Z < z) = 0.975$.

.

Solution

(a) From the standard normal table, we have

$$P(Z < 1.25) = \Phi(1.25) \approx 0.8944.$$

(b) From the standard normal table, we have

$$P(Z < 0.85) = \Phi(0.85) \approx 0.8023$$

Therefore,

$$P(Z > 0.85) = 1 - \Phi(0.85) \approx 1 - 0.8023 = 0.1977.$$

(c) From the standard normal table, we have

$$P(Z < 0.5) = \Phi(0.5) \approx 0.6915, \quad P(Z < -1.2) = \Phi(-1.2) \approx 0.1151$$

Therefore,

$$P(-1.2 < Z < 0.5) = \Phi(0.5) - \Phi(-1.2) \approx 0.6915 - 0.1151 = 0.5764.$$

DRAFT

(d) We have,

$$\begin{aligned} P(|Z| < 2) &= P(-2 < Z < 2) \\ &= \Phi(2) - \Phi(-2) \\ &\approx 0.9772 - 0.0228 = 0.9544. \end{aligned}$$

Alternatively,

$$\begin{aligned} P(|Z| < 2) &= 1 - P(|Z| \geq 2) \\ &= 1 - [P(Z \geq 2) + P(Z \leq -2)] \\ &= 1 - [1 - \Phi(2) + \Phi(-2)] \\ &= \Phi(2) - \Phi(-2) \\ &\approx 0.9772 - 0.0228 = 0.9544. \end{aligned}$$

(e) From standard normal distribution table, we have

$$\Phi(1.96) = 0.975$$

Since the problem requires $P(Z < z) = 0.975$. it follows that

$$z = 1.96.$$

Remark 7.4.1. *The absolute value probability inequalities can always be rewritten in terms of standard inequalities. This is particularly useful when working with continuous random variables such as the normal distribution.*

- *For **inside intervals**, i.e., $|X - a| < b$, the event represents values of X within b units of a, and can be rewritten as:*

$$|X - a| < b \iff a - b < X < a + b.$$

- *For **outside intervals**, i.e., $|X - a| \geq b$, the event represents values of X at least b units away from a, and can be rewritten as:*

$$|X - a| \geq b \iff X \geq a + b \text{ or } X \leq a - b.$$

7.4.5 Relationship between X and Z (Z-score)

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normal random variable. The **Z-score** of X is defined as

$$Z = \frac{X - \mu}{\sigma}, \quad Z \sim \mathcal{N}(0, 1).$$

This measures how many standard deviations X is above or below the mean μ.

Conversely, any Z-score can be converted back to the original variable X by:

$$X = \mu + Z\sigma.$$

The Z-score allows us to use standard normal tables to compute probabilities for any normal variable X.

DRAFT

Finding the Probability $P(a \leq X \leq b)$

The probability $P(a \leq X \leq b)$ for a continuous random variable X with probability density function $f(x)$ is defined as the area under the curve of the PDF between a and b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

as illustrated in Figure 7.12. Alternatively, this probability can be expressed in terms of the cumulative distribution function $F(x)$ of X as:

$$P(a \leq X \leq b) = F(b) - F(a).$$

In the case of the normal distribution, direct computation of the CDF $F(x)$ is not straightforward because it does not have a closed-form expression. However, we can make use of the standard normal distribution to simplify the calculation. The standard normal distribution, denoted by $Z \sim \mathcal{N}(0,1)$, has a tabulated CDF $\Phi(z)$, which allows us to compute probabilities more easily.

To compute the probability $P(a \leq X \leq b)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$, follow these steps:

(i). Since

$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1),$$

we are transforming our general normal random variable X into a standard normal random variable, usually denoted by Z. This transformation is done using the **Z-score** formula:

$$Z = \frac{X-\mu}{\sigma}$$

which gives:

$$P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right).$$

(ii). Use the standard normal cumulative distribution function $\Phi(z)$ to compute the probability:

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(Z \leq \frac{a-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

DRAFT

(iii). Substitute numerical values and evaluate the result using standard normal tables or statistical software.

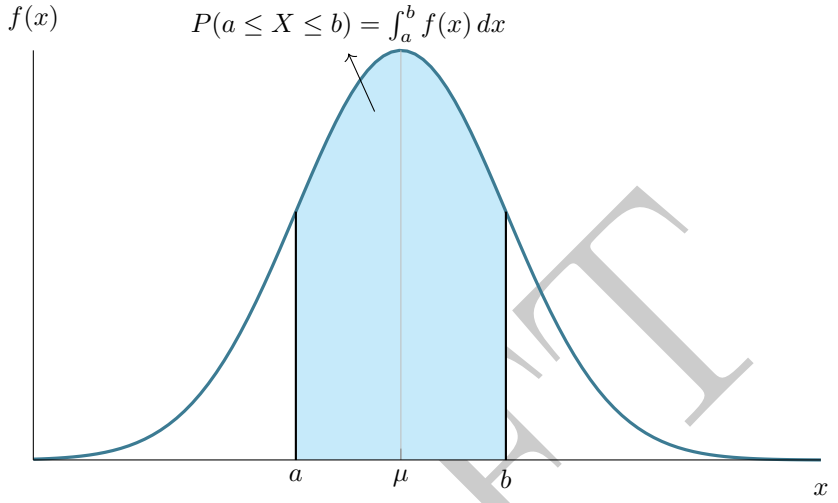


Figure 7.12: The area under the probability density function $f(x)$ between a and b.

Finding the Probability $P(a \leq X \leq b)$**:** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the probability $P(a \leq X \leq b)$ can be computed as

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

Problem 7.8. *The scores on a standardized test are normally distributed with a mean (μ) of 75 and a standard deviation (σ) of 8.*

(a). What is the probability that a randomly selected student scored above 85 on the test?

(b). What is the probability that a randomly selected student scored between 60 and 70 on the test?

(c). What is the probability that a randomly selected student scored at most 90 on the test?

DRAFT

Solution

Let X be the score of a randomly selected student on the standardized test. We are given that X is normally distributed, where $X \sim \mathcal{N}(75, 8^2)$.

We will use the Z-score formula $Z = \frac{X-\mu}{\sigma}$ and the standard normal cumulative distribution function $\Phi(z)$ to solve each part.

(a) Probability that a student scored above 85: We want to find $P(X > 85)$.

$$\begin{aligned} P(X > 85) &= P\left(\frac{X-\mu}{\sigma} > \frac{85-75}{8}\right) \\ &= P(Z > 1.25) \\ &= 1 - \Phi(1.25) \\ &\approx 1 - 0.8944 \\ &= 0.1056 \end{aligned}$$

The probability that a randomly selected student scored above 85 is approximately 0.1056.

(b) Probability that a student scored between 60 and 70: We want to find $P(60 \leq X \leq 70)$. Hence,

$$\begin{aligned} P(60 \leq X \leq 70) &= P\left(\frac{60-75}{8} \leq \frac{X-\mu}{\sigma} \leq \frac{70-75}{8}\right) \\ &= P(-1.88 \leq Z \leq -0.63) \\ &= \Phi(-0.63) - \Phi(-1.88) \\ &\approx 0.2643 - 0.0301 \\ &= 0.2342 \end{aligned}$$

The probability that a randomly selected student scored between 60 and 70 is approximately 0.2342.

(c) Probability that a student scored at most 90: We want to find $P(X \leq 90)$. Thus,

$$\begin{aligned} P(X \leq 90) &= P\left(\frac{X-\mu}{\sigma} \leq \frac{90-75}{8}\right) \\ &= P(Z \leq 1.88) \\ &= \Phi(1.88) \\ &\approx 0.9699 \end{aligned}$$

DRAFT

The probability that a randomly selected student scored at most 90 is approximately 0.9699.

Problem 7.9. *Suppose* $X \sim \mathcal{N}(10, 4)$. *Find* $P(8 \leq X \leq 12)$.

Solution
Since, $X \sim \mathcal{N}(10, 4)$, then

$$\begin{aligned} P(8 \leq X \leq 12) &= P\left(\frac{8-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{12-\mu}{\sigma}\right) \\ &= P\left(\frac{8-10}{2} \leq Z \leq \frac{12-10}{2}\right) \\ &= P\left(-1 \leq Z \leq 1\right) \\ &= \Phi(1) - \Phi(-1) \\ &\approx 0.8413 - 0.1587 = 0.6826 \end{aligned}$$

Thus, the probability that X lies between 8 and 12 is approximately 0.6826.

Problem 7.10. *Suppose that* $Y \sim \mathcal{N}(10, 16)$. *What is the value of* $P(|Y-10| \geq 12)$*?*

Solution

Given $Y \sim \mathcal{N}(10, 16)$, we know that the mean $\mu = 10$ and standard deviation $\sigma = 4$. We need to compute $P(|Y - 10| \geq 12)$. This is equivalent to the following:

$$\begin{aligned} P(|Y-10| \geq 12) &= P(Y - 10 \geq 12 \text{ or } -(Y-10) \geq 12) \\ &= P(Y \geq 22 \text{ or } Y \leq -2) \\ &= P(Y \geq 22) + P(Y \leq -2) \\ &= P\left(\frac{Y-\mu}{\sigma} \geq \frac{22-10}{4}\right) + P\left(\frac{Y-\mu}{\sigma} \leq \frac{-2-10}{4}\right) \\ &= P\left(Z \geq 3\right) + P\left(Z \leq -3\right) \\ &= [1 - \Phi(3)] + \Phi(-3) \\ &\approx [1 - 0.99865] + 0.00135 = 0.0027 \end{aligned}$$

□

Problem 7.11. *Assume the heights of adult males in a certain population are normally distributed with a mean of* 70 *inches and a standard deviation of* 3 *inches. What is the probability that a randomly selected adult male has a height between* 67 *and* 73 *inches?*

DRAFT

Solution

Let X denote the height of a randomly selected adult male, where $X \sim \mathcal{N}(70, 3^2)$.

To find the probability that a randomly selected adult male has a height between 67 and 73 inches, we standardize the values and use the standard normal distribution. Therefore,

$$\begin{aligned} P(67 \leq X \leq 73) &= P\left(\frac{67-70}{3} \leq Z \leq \frac{73-70}{3}\right) \\ &= P\left(-1 \leq Z \leq 1\right) \\ &= \Phi(1) - \Phi(-1) \\ &\approx 0.8413 - 0.1587 = 0.6826 \end{aligned}$$

Thus, approximately 68.26% of adult males have heights between 67 and 73 inches.

□

Problem 7.12. *What is the value of a for which $P(Z \geq a) = 0.72$, where $Z \sim \mathcal{N}(0, 1)$?*

Solution

We are given

$$P(Z \geq a) = 0.72.$$

First convert this to a left-tail probability:

$$P(Z < a) = 1 - 0.72 = 0.28.$$

From the standard normal distribution table, we have

$$\Phi(-0.58) \approx 0.28.$$

Since the problem requires

$$P(Z < a) = 0.28,$$

it follows that

$$a \approx -0.58.$$

Therefore, the value of a for which $P(Z \geq a) = 0.72$ is approximately -0.5828. □

DRAFT

Python Code Solution

The following Python code uses the `scipy.stats` library to calculate the value of a.

```
from scipy.stats import norm

# Given: P(Z >= a) = 0.72
# We know that P(Z >= a) = 1 - P(Z < a)
# So, P(Z < a) = 1 - P(Z >= a)

probability_less_than_a = 1 - 0.72

# Use the Percent Point Function (ppf), which is the inverse of the CDF,
# to find the Z-score (value of 'a') corresponding to the cumulative probability.
a_value = norm.ppf(probability_less_than_a)

print(f"Given P(Z >= a) = 0.72")
print(f"Therefore, P(Z < a) = {probability_less_than_a:.4f}")
print(f"The value of a is approximately: {a_value:.4f}")
```

Problem 7.13. *Given a standard normal distribution, find the value of k such that*

(a). $P(Z > k) = 0.3015$

(b). $P(k < Z < -0.18) = 0.4197$.

Solution

For a standard normal distribution, let $\Phi(z)$ denote the cumulative distribution function (CDF), i.e., $\Phi(z) = P(Z \leq z)$.

(a) Find the value of k such that $P(Z > k) = 0.3015$

We are given

$$P(Z > k) = 0.3015.$$

Convert this to a left–tail probability:

$$P(Z < k) = 1 - 0.3015 = 0.6985.$$

From the standard normal distribution table, we have

$$\Phi(0.52) \approx 0.6985.$$

Since the problem requires

$$P(Z < k) = 0.6985,$$

DRAFT

it follows that

$$k \approx 0.52.$$

(b) Find the value of k such that $P(k < Z < -0.18) = 0.4197$

We are given

$$P(k < Z < -0.18) = 0.4197.$$

First compute the left–tail value at -0.18:

$$P(Z < -0.18) = \Phi(-0.18) \approx 0.4286.$$

Using the probability statement,

$$P(k < Z < -0.18) = \Phi(-0.18) - \Phi(k) = 0.4197.$$

Therefore,

$$\Phi(k) \approx 0.4286 - 0.4197 = 0.0089.$$

From the standard normal table,

$$\Phi(-2.37) \approx 0.0089.$$

Since the problem requires

$$\Phi(k) \approx 0.0089,$$

it follows that

$$k \approx -2.37.$$

Python Code Solution

The following Python code uses the `scipy.stats` library to calculate the values of k.

```
from scipy.stats import norm

# --- Part (a) ---
# Given: P(Z > k) = 0.3015
# This is equivalent to P(Z <= k) = 1 - 0.3015
p_le_k_a = 1 - 0.3015
k_a = norm.ppf(p_le_k_a)

print(f"For part (a):")
print(f"P(Z <= k) = {p_le_k_a:.4f}")
print(f"The value of k is approximately: {k_a:.4f}")
print("-" * 30)

# --- Part (b) ---
# Given: P(k < Z < -0.18) = 0.4197
```

DRAFT

```
# We know that P(k < Z < -0.18) = P(Z < -0.18) - P(Z < k)

# Step 1: Find P(Z < -0.18) using the CDF
p_lt_minus_0_18 = norm.cdf(-0.18)

# Step 2: Use the given equation to find P(Z < k)
# 0.4197 = p_lt_minus_0_18 - P(Z < k)
# P(Z < k) = p_lt_minus_0_18 - 0.4197
p_lt_k_b = p_lt_minus_0_18 - 0.4197

# Step 3: Find k using the inverse CDF (ppf)
k_b = norm.ppf(p_lt_k_b)

print(f"For part (b):")
print(f"P(Z < -0.18) = {p_lt_minus_0_18:.4f}")
print(f"P(Z < k) = {p_lt_k_b:.4f}")
print(f"The value of k is approximately: {k_b:.4f}")

```

□

Problem 7.14. *Suppose the heights of a large population of adult males in a certain country follow a normal distribution with a mean height of 175 centimeters and a standard deviation of 6 centimeters. Calculate the height below which the shortest 10% of adult males in this population fall.*

Solution
We are asked to find the height below which the lowest 10% of adult males fall. Let X denote the height, where

$$X \sim \mathcal{N}(175, 6^2).$$

The shortest 10% corresponds to the 0.10 quantile:

$$P(X < x_{0.10}) = 0.10.$$

First find the corresponding standard normal value:

$$z_{0.10} = -1.28,$$

since $\Phi(-1.28) = 0.10$.

Now convert this to the original distribution using

$$x_{0.10} = \mu + z\sigma.$$

Thus,

$$x_{0.10} = 175 + (-1.28)(6) = 175 - 7.68 = 167.32.$$

Therefore, the shortest 10% of adult males have heights below 167.32 cm.

□

DRAFT

Problem 7.15. *The Wall Street Journal Interactive Edition spend average of 27 hours per week using the computer at work. Assume the normal distribution applies and that the standard deviation is 8 hours.*

(a). What is the probability a randomly selected subscriber spends less than 10 hours using the computer at work?

(b). What percentage of the subscribers spends more than 35 hours per week using the computer at work?

(c). A person is classified as a heavy user if he or she is in the upper 20% in terms of hours of usage. How many hours must a subscriber use the computer in order to be classified as a heavy user?

Solution
Let X be the number of hours a randomly selected subscriber spends using the computer at work per week. Assume X follows a normal distribution:

$$X \sim \mathcal{N}(27, 8^2)$$

where $\mu = 27$ hours and $\sigma = 8$ hours.

(a). To find the probability that a randomly selected subscriber spends less than 10 hours using the computer at work, we need to find $P(X < 10)$.

$$\begin{aligned} P(X < 10) &= P\left(\frac{X-\mu}{\sigma} < \frac{10-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{10-27}{8}\right) \\ &= P(Z < -2.125) \\ &= \Phi(-2.125) \\ &\approx 0.0168 \end{aligned}$$

Thus, the probability that a randomly selected subscriber spends less than 10 hours using the computer at work is approximately 0.0168.

(b). To find the percentage of subscribers who spend more than 35 hours per week using the computer at work, we need to find $P(X > 35)$.

$$\begin{aligned} P(X > 35) &= P\left(\frac{X-\mu}{\sigma} > \frac{35-\mu}{\sigma}\right) \\ &= P(Z > 1) \\ &= 1 - \Phi(1) \\ &\approx 1 - 0.8413 \\ &= 0.1587 \end{aligned}$$

Thus, approximately 15.87% of the subscribers spend more than 35 hours per week using the computer at work.

DRAFT

(c). A person is classified as a heavy user if he or she is in the upper 20% in terms of hours of usage. This means we are looking for a value a such that

$$P(X > a) = 0.20.$$

If $P(X > a) = 0.20$, then $P(X \leq a) = 1 - 0.20 = 0.80$. We need to find the Z-score $z_{0.80}$ such that $P(Z \leq z_{0.80}) = 0.80$.

From the standard normal table, we find that $\Phi(z_{0.80}) \approx 0.80$ corresponds to $z_{0.80} \approx 0.84$. Now, we use the formula $a = \mu + z_{0.80}\sigma$ to find the corresponding value of a:

$$\begin{aligned} a &= 27 + (0.84)(8) \\ &= 27 + 6.72 \\ &= 33.72 \end{aligned}$$

Thus, a subscriber must use the computer for approximately 33.72 hours per week to be classified as a heavy user.

□

7.4.6 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental theorem in probability theory and statistics. It states that the distribution of the sum (or average) of a large number of independent, identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution of the variables. This theorem underpins many statistical methods and justifies the use of the normal distribution in inferential statistics.

Theorem 7.7 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\bar{X}_n$ denote the sample mean:*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1} X_i$$

Then, as n approaches infinity, the distribution of the standardized sample mean approaches the standard normal distribution:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

or equivalently,

$$\bar{X}_n \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n)$$

where $\xrightarrow{d}$ denotes convergence in distribution.

DRAFT

Problem 7.16. *Suppose we have a fair six-sided die. The outcome of a single roll, X, is a random variable with a uniform distribution on the integers $\{1, 2, 3, 4, 5, 6\}$. The mean of a single roll is $\mu = (1+2+3+4+5+6)/6 = 3.5$. The variance of a single roll is $\sigma^2 = E[X^2] - (E[X])^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - (3.5)^2 = \frac{91}{6} - 12.25 \approx 15.1667 - 12.25 = 2.9167$. So, the standard deviation is $\sigma = \sqrt{2.9167} \approx 1.7078$.*

If we roll the die 40 times and calculate the average of these 40 rolls, what is the approximate probability that the average of these 40 rolls is less than 3?

Solution

Let $X_1, X_2, \ldots, X_{40}$ be the outcomes of 40 independent rolls of a fair six-sided die. These are i.i.d. random variables with $\mu = 3.5$ and $\sigma = 1.7078$. Let $\bar{X}_{40}$ be the sample mean of these 40 rolls.

According to the Central Limit Theorem, since the sample size $n = 40$ is sufficiently large, the distribution of the sample mean $\bar{X}_{40}$ can be approximated by a normal distribution with:

- Mean of the sample mean: $E[\bar{X}_{40}] = \mu = 3.5$
- Variance of the sample mean: $Var(\bar{X}_{40}) = \frac{\sigma^2}{n} = \frac{2.9167}{40} \approx 0.0729$
- Standard deviation of the sample mean (standard error): $\sigma_{\bar{X}_{40}} = \frac{\sigma}{\sqrt{n}} = \frac{1.7078}{\sqrt{40}} \approx \frac{1.7078}{6.3246} \approx 0.2700$

So, we can approximate $\bar{X}_{40} \sim \mathcal{N}(3.5, 0.0729)$.

We want to find the approximate probability that the average of these 40 rolls is less than 3, i.e., $P(\bar{X}_{40} < 3)$. We standardize $\bar{X}_{40}$ to a standard normal variable Z:

$$\begin{aligned} P(\bar{X}_{40} < 3) &= P\left(\frac{\bar{X}_{40} - \mu}{\sigma/\sqrt{n}} < \frac{3 - 3.5}{0.2700}\right) \\ &= P\left(Z < \frac{-0.5}{0.2700}\right) \\ &= P(Z < -1.85) \\ &= \Phi(-1.85) \\ &\approx 0.0322 \end{aligned}$$

Thus, the approximate probability that the average of 40 die rolls is less than 3 is 0.0322. This example demonstrates how the distribution of sample means tends towards a normal distribution, even when the original population distribution (a single die roll) is uniform.

Problem 7.17. *A researcher is studying the systolic blood pressure levels in a population of adults. It is known that the systolic blood pressure levels are normally distributed with a mean (μ) of 120 mmHg and a standard deviation (σ) of 15 mmHg.*

DRAFT

(a). What proportion of the population has systolic blood pressure levels between 110 mmHg and 130 mmHg?

(b). What is the probability that a randomly selected individual from this population has a systolic blood pressure level above 140 mmHg?

(c). If the researcher takes a random sample of 25 adults, what is the probability that the sample mean systolic blood pressure is less than 115 mmHg?

Solution (a). Proportion of the Population between 110 mmHg and 130 mmHg
We need to find $P(110 \leq X \leq 130)$ where X is the systolic blood pressure level.

$$\begin{aligned} P(110 \leq X \leq 130) &= P\left(\frac{110-120}{15} \leq \frac{X-\mu}{\sigma} \leq \frac{130-120}{15}\right) \\ &= P(-0.67 \leq Z \leq 0.67) \\ &= P(Z \leq 0.67) - P(Z \leq -0.67) \\ &= \Phi(0.67) - \Phi(-0.67) \\ &= 0.7486 - 0.2514 \\ &= 0.4972 \end{aligned}$$

So, approximately 49.72% of the population has systolic blood pressure levels between 110 mmHg and 130 mmHg.

(b). Probability of Blood Pressure Above 140 mmHg
We need to find $P(X > 140)$.

$$\begin{aligned} P(X > 140) &= P\left(\frac{X-\mu}{\sigma} > \frac{140-120}{15}\right) \\ &= 1 - P(Z \leq 1.33) \\ &\approx 1 - 0.9082 = 0.0918 \end{aligned}$$

So, the probability that a randomly selected individual has a systolic blood pressure level above 140 mmHg is approximately 0.0918 or 9.18%.

(c). Probability of Sample Mean Less Than 115 mmHg
For a sample of $n = 25$, the distribution of the sample mean $\bar{X}$ is normally distributed with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
Given:

$$\mu_{\bar{X}} = 120, \quad \sigma_{\bar{X}} = \frac{15}{\sqrt{25}} = 3$$

DRAFT

Hence,

$$\begin{aligned} P(\bar{X} < 115) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{115 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(Z \leq \frac{115 - 120}{3}\right) \\ &= P(Z \leq -1.67) = \Phi(-1.67) \approx 0.0475 \end{aligned}$$

So, the probability that the sample mean systolic blood pressure of 25 adults is less than 115 mmHg is approximately 0.0475 or 4.75%.

Problem 7.18. *A study is conducted to measure the cholesterol levels in a population of adults. It is known that the cholesterol levels are normally distributed with a mean (μ) of 200 mg/dL and a standard deviation (σ) of 25 mg/dL.*

(a). What percentage of the population has cholesterol levels between 175 mg/dL and 225 mg/dL?

(b). What is the probability that a randomly selected individual has a cholesterol level below 180 mg/dL?

(c). If a sample of 36 adults is taken, what is the probability that the sample mean cholesterol level is greater than 210 mg/dL?

Solution

(a). Percentage of the Population between 175 mg/dL and 225 mg/dL

Let X is the cholesterol level. Then,

$$\begin{aligned} P(175 \leq X \leq 225) &= P\left(\frac{175 - 200}{25} \leq \frac{X - \mu}{\sigma} \leq \frac{225 - 200}{25}\right) \\ &= P\left(-1 \leq Z \leq 1\right) \\ &= \Phi(1) - \Phi(-1) \\ &\approx 0.8413 - 0.1587 = 0.6826 \end{aligned}$$

So, approximately 68.26% of the population has cholesterol levels between 175 mg/dL and 225 mg/dL.

(b). Probability of Cholesterol Level Below 180 mg/dL

$$\begin{aligned} P(X < 180) &= P\left(\frac{X - \mu}{\sigma} < \frac{180 - 200}{25}\right) \\ &= P(Z \leq -0.8) \\ &= \Phi(-0.8) \approx 0.2119 \end{aligned}$$

So, the probability that a randomly selected individual has a cholesterol level below 180 mg/dL is approximately 0.2119 or 21.19%.

DRAFT

(c). Probability of Sample Mean Greater Than 210 mg/dL

For a sample of $n = 36$, the distribution of the sample mean $\bar{X}$ is normally distributed with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Given:

$$\mu_{\bar{X}} = 200, \quad \sigma_{\bar{X}} = \frac{25}{\sqrt{36}} = \frac{25}{6} \approx 4.17$$

We need to find $P(\bar{X} > 210)$. Thus,

$$\begin{aligned} P(\bar{X} > 210) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{210 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P(Z > 2.40) = 1 - \Phi(2.40) \\ &\approx 1 - 0.9918 = 0.0082 \end{aligned}$$

So, the probability that the sample mean cholesterol level of 36 adults is greater than 210 mg/dL is approximately 0.0082 or 0.82%.

7.4.7 Approximating the Binomial Distribution with the Normal Distribution

The **binomial distribution** is one of the most important discrete probability distributions which is described in Section 6.3. However, for large values of n, computing exact binomial probabilities becomes computationally intensive. In such cases, the **normal distribution** provides a very useful approximation.

The Binomial Distribution

A random variable $X \sim \text{Bin}(n, p)$ counts the number of successes in n independent Bernoulli trials, each with probability of success p. The mean and variance of X are given by

$$\mu = E[X] = np, \qquad \sigma^2 = \text{Var}(X) = np(1-p).$$

Conditions for Normal Approximation

Since the binomial distribution is discrete and the normal distribution is continuous, the approximation works well only under certain conditions. A commonly used rule of thumb is:

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5.$$

These conditions ensure that the binomial distribution is sufficiently symmetric and bell-shaped.

DRAFT

The Normal Approximation

Under the above conditions, the binomial distribution can be approximated as

$$X \sim \text{Bin}(n, p) \quad \approx \quad Y \sim N\big(np,\, np(1-p)\big).$$

Thus, probabilities involving X can be approximated using the standard normal variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1).$$

Continuity Correction

Because the binomial distribution is discrete and the normal distribution is continuous, a **continuity correction** should be applied for better accuracy. The correction involves shifting by 0.5 in the appropriate direction:

- $P(X = k) \approx P(k - 0.5 < Y < k + 0.5)$
- $P(X \leq k) \approx P(Y \leq k + 0.5)$
- $P(X \geq k) \approx P(Y \geq k - 0.5)$

This correction significantly improves the quality of the approximation.

Problem 7.19. *Suppose $X \sim Bin(n = 50, p = 0.4)$. We wish to approximate $P(X \leq 18)$.*

1. **Check Conditions:** The conditions for a normal approximation to the binomial distribution are met as both np and $n(1-p)$ are greater than 5.
   - $np = 50(0.4) = 20 \geq 5$
   - $n(1-p) = 50(0.6) = 30 \geq 5$
2. **Define the Normal Distribution:** We calculate the mean (μ) and standard deviation (σ) for the approximating normal distribution.
   - Mean: $\mu = np = 50(0.4) = 20$
   - Standard Deviation: $\sigma = \sqrt{np(1-p)} = \sqrt{50(0.4)(0.6)} = \sqrt{12} \approx 3.464$
3. **Apply Continuity Correction:** To approximate the discrete probability $P(X \leq 18)$ using a continuous normal distribution, we apply a continuity correction. The value $X = 18$ corresponds to the continuous interval up to 18.5.

$$P(X \leq 18) \approx P(Y \leq 18.5)$$

DRAFT

4. **Calculate the Z-score and Probability:** We standardize the value using the Z-score formula:

$$Z = \frac{Y - \mu}{\sigma} = \frac{18.5 - 20}{3.464} = \frac{-1.5}{3.464} \approx -0.433$$

Using a standard normal (Z) table or calculator, the probability is:

$$P(Y \leq 18.5) = P(Z \leq -0.433) = \Phi(-0.433) \approx 0.3326$$

Thus, the normal approximation yields an approximate probability of 0.333.

Problem 7.20. *A multiple-choice quiz has 100 questions, with each question having 4 possible answers. A student guesses the answer to every question. What is the approximate probability that the student gets at least 30 questions correct?*

Solution

1. **Check Conditions:**
   - Number of trials, $n = 100$.
   - Probability of success, $p = 1/4 = 0.25$.
   - Check the conditions for normal approximation:
     - $np = 100(0.25) = 25 \geq 5$
     - $n(1-p) = 100(0.75) = 75 \geq 5$

   Since both conditions are met, the normal approximation is appropriate.

2. **Define the Normal Distribution:**
   - Mean: $\mu = np = 100(0.25) = 25$
   - Standard Deviation: $\sigma = \sqrt{np(1-p)} = \sqrt{100(0.25)(0.75)} = \sqrt{18.75} \approx 4.33$

3. **Apply Continuity Correction:** The binomial probability of "at least 30" questions correct, $P(X \geq 30)$, is approximated by the normal probability $P(X > 29.5)$.

4. **Calculate the Z-score:** $Z = \frac{X-\mu}{\sigma} = \frac{29.5-25}{4.33} = \frac{4.5}{4.33} \approx 1.04$

5. **Find the Probability:** Using a Z-table, the area to the left of $Z = 1.04$ is approximately 0.8508. The area to the right is the probability we are looking for:

$$P(X > 29.5) = P(Z > 1.04) = 1 - \Phi(1.04) = 1 - 0.8508 = 0.1492$$

The approximate probability that the student gets at least 30 questions correct is 0.1492.

DRAFT

7.4.8 Python Code for Normal Distribution Characteristics

The Normal (Gaussian) distribution is fundamental in statistics and is used to model continuous random variables. Below is Python code that demonstrates how to compute various characteristics of the Normal distribution.

Python Code

```
import numpy as np
from scipy.stats import norm

# Define the parameters
mu = 0        # Mean
sigma = 1     # Standard deviation

# Normal distribution
dist = norm(loc=mu, scale=sigma)

# 1. Probability Density Function (PDF)
x_values = np.linspace(-5, 5, 100)  # Values for the
  random variable
pdf_values = dist.pdf(x_values)
print("PDF values for x from -5 to 5:", pdf_values)

# 2. Cumulative Distribution Function (CDF)
cdf_values = dist.cdf(x_values)
print("CDF values for x from -5 to 5:", cdf_values)

# 3. Mean (Expected Value)
mean = dist.mean()
print("Mean (Expected Value):", mean)

# 4. Variance
variance = dist.var()
print("Variance:", variance)

# 5. Standard Deviation
std_dev = dist.std()
print("Standard Deviation:", std_dev)

# 6. Quantiles
quantiles = dist.ppf([0.25, 0.5, 0.75])  # 25th, 50th (
  median), and 75th percentiles
print("Quantiles at 0.25, 0.5, and 0.75:", quantiles)

# 7. Percentiles
```

```
percentiles = dist.ppf([0.1, 0.9])  # 10th and 90th
  percentiles
print("Percentiles at 0.1 and 0.9:", percentiles)

# 8. Moment Generating Function (MGF)
def mgf(t, mu, sigma):
return np.exp(mu * t + 0.5 * (sigma**2) * t**2)

# MGF values at t = 0 and t = 1
mgf_values = [mgf(t, mu, sigma) for t in [0, 1]]
print("MGF values for t = 0 and t = 1:", mgf_values)
```

Explanations

- **Probability Density Function (PDF):**
  - The PDF provides the likelihood of each value. Compute these values using `dist.pdf(x_values)`.
- **Cumulative Distribution Function (CDF):**
  - The CDF gives the probability that the variable takes on a value less than or equal to a given value. Compute this using `dist.cdf(x_values)`.
- **Mean (Expected Value):**
  - The mean (expected value) is μ. Compute this using `dist.mean()`.
- **Variance:**
  - The variance is σ^2. Compute this using `dist.var()`.
- **Standard Deviation:**
  - The standard deviation is σ. Compute this using `dist.std()`.
- **Quantiles:**
  - Quantiles are values at specific probabilities. Compute these using `dist.ppf()` for the desired quantile probabilities.
- **Percentiles:**
  - Percentiles are specific types of quantiles. Compute these using `dist.ppf()` for the desired percentile probabilities.
- **Moment Generating Function (MGF):**

DRAFT

- The MGF for a Normal distribution is given by $M_X(t) = \exp(\mu \cdot t + 0.5 \cdot \sigma^2 \cdot t^2)$. Define a function `mgf(t, mu, sigma)` to compute MGF values.

7.4.9 Exercises

1. Given a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, answer the following questions:
   (a) If $\mu = 10$ and $\sigma = 2$, what is the probability that X is less than 8?
   (b) Find the probability that X lies between 8 and 12.
   (c) Determine the value a such that $P(X \leq a) = 0.975$.
2. The standard normal distribution is a special case of the normal distribution where $\mu = 0$ and $\sigma = 1$. Use the standard normal distribution table (z-table) to answer the following:
   (a) Find $P(Z \leq 1.645)$ for $Z \sim \mathcal{N}(0, 1)$.
   (b) Determine $P(-1.96 \leq Z \leq 1.96)$.
   (c) Calculate the value z such that $P(Z \geq z) = 0.05$.
3. Given $X \sim \mathcal{N}(20, 25)$, transform X to the standard normal distribution Z and solve the following:
   (a) Find the probability that X is less than 15.
   (b) Determine the probability that X lies between 18 and 22.
   (c) Find the value x such that $P(X \leq x) = 0.90$.
4. Consider a random variable Y representing the strength of a material, which follows a normal distribution with mean 500 MPa and standard deviation 50 MPa.
   (a) What is the probability that the strength is between 450 MPa and 550 MPa?
   (b) What is the 95th percentile of the strength?
   (c) Calculate the variance of the strength.
   (d) What proportion of material samples have a strength greater than 600 MPa?
5. Suppose $X \sim \mathcal{N}(50, 25)$. What is $P(45 \leq X \leq 55)$?
6. Answer the following application-based questions:
   (a) The heights of adult men in a certain population are normally distributed with a mean of 175 cm and a standard deviation of 10 cm. What percentage of men are taller than 190 cm?

DRAFT

(b) A factory produces light bulbs with lifetimes that are normally distributed with a mean of 1200 hours and a standard deviation of 100 hours. What is the probability that a randomly selected light bulb lasts between 1100 and 1300 hours?

(c) The scores on a standardized test are normally distributed with a mean of 500 and a standard deviation of 100. What is the minimum score needed to be in the top 5% of test-takers?

7. Given a population with mean $\mu = 60$ and standard deviation $\sigma = 15$:

   (a) If a sample of size 50 is taken, what is the expected value and standard deviation of the sample mean?

   (b) Using the Central Limit Theorem, find the probability that the sample mean is less than 58.

   (c) Calculate the probability that the sample mean lies between 59 and 62.

8. A study measures the cholesterol levels (in mg/dL) of a group of patients, which are found to follow a normal distribution with a mean of 200 mg/dL and a standard deviation of 20 mg/dL.

   (a) What is the probability that a randomly selected patient has a cholesterol level between 180 mg/dL and 220 mg/dL?

   (b) What is the 95th percentile of the cholesterol levels?

   (c) Calculate the mean and standard deviation of sample mean of the cholesterol levels.

   (d) If a cholesterol level above 240 mg/dL is considered high, what proportion of the patients have high cholesterol levels?

9. Consider a population where the heights of adult women are approximately normally distributed with a mean of 65 inches and a standard deviation of 4 inches.

   (a) Using Chebyshev's Inequality, find the minimum percentage of women whose heights are within 10 inches of the mean height.

   (b) Suppose you want to ensure that at least 95% of the women fall within a certain number of standard deviations from the mean. Using Chebyshev's Inequality, determine the minimum number of standard deviations required for this guarantee.

   (c) In the same population, if the height of a randomly selected woman is 70 inches, what is the probability that her height deviates from the mean by at least 5 inches, according to Chebyshev's Inequality?

10. The systolic blood pressure of a certain population is normally distributed with a mean of 120 mmHg and a standard deviation of 15 mmHg.

DRAFT

(a) What is the probability that a randomly selected person has a systolic blood pressure less than 110 mmHg?

(b) What is the probability that a randomly selected person has a systolic blood pressure between 110 mmHg and 130 mmHg?

(c) Find the 95th percentile of the systolic blood pressure distribution.

11. Suppose X and Y are independent random variables representing the weights (in kg) of two different species of fish in a lake. X follows a normal distribution with mean 2 kg and variance 0.25 kg^2, and Y follows a normal distribution with mean 3 kg and variance 0.36 kg^2.

    (a) What is the probability that a fish of the first species weighs more than 2.5 kg?

    (b) What is the probability that a fish of the second species weighs between 2.5 kg and 3.5 kg?

    (c) What is the expected value of the difference in weights $D = X - Y$?

    (d) What is the variance of the difference in weights D?

    ,

12. Consider a sample mean $\bar{X}$ from a normal distribution with population mean μ and variance σ^2. Assume the sample size is $n = 36$.

    (a) If $\mu = 50$ and $\sigma = 12$, find the probability that the sample mean is greater than 52.

    (b) Calculate the probability that the sample mean lies between 48 and 51.

    (c) Determine the value of $\bar{x}$ such that $P(\bar{X} \leq \bar{x}) = 0.95$.

7.5 Concluding Remarks

In this chapter, we have examined several essential continuous probability distributions, highlighting their definitions, properties, and applications. The Uniform distribution provided a foundation for understanding equal likelihood over a range of values, while the Exponential distribution illustrated the modeling of time intervals between events in a Poisson process.

The Normal distribution, with its profound significance in statistical theory and practice, was explored in detail. We discussed its properties, the concept of the Standard Normal distribution, methods for finding areas under the normal curve, and the Central Limit Theorem, which underscores the Normal distribution's ubiquitous presence in statistical analysis.

DRAFT

Understanding these continuous distributions equips us with powerful tools for modeling and analyzing data in numerous disciplines. By mastering these concepts, we can better interpret real-world phenomena, make informed decisions, and contribute to advancements in various fields.

As we continue our journey through probability and statistics, the knowledge gained from studying continuous distributions will serve as a crucial foundation for more complex analyses and applications.

7.6 Chapter Exercises

1. A random variable X follows a continuous uniform distribution between 2 and 5. Calculate the probability that X is less than 3.

2. If the length of a rod is uniformly distributed between 10 and 20 cm, what is the probability that a randomly selected rod is longer than 15 cm?

3. The time between arrivals of customers at a coffee shop follows an exponential distribution with a mean of 5 minutes. What is the probability that the next customer will arrive within 3 minutes?

4. A radioactive substance has a half-life of 10 years. What is the probability that a sample will decay in less than 5 years?

5. A set of test scores is normally distributed with a mean of 70 and a standard deviation of 10. What is the probability that a randomly selected score is greater than 85?

6. In a factory, the weight of bags of flour is normally distributed with a mean of 50 kg and a standard deviation of 2 kg. What percentage of bags weigh between 48 kg and 52 kg?

7. A company's delivery times follow a normal distribution with a mean of 30 minutes and a standard deviation of 5 minutes. What is the probability that a delivery takes longer than 40 minutes?

8. Compare the probabilities of an event occurring within a specified time frame using both the exponential distribution (with a mean of 6 minutes) and the continuous uniform distribution (from 0 to 12 minutes).

9. A car rental service finds that the time a customer spends renting a car follows a normal distribution with a mean of 4 days and a standard deviation of 1.5 days. What is the probability that a customer rents a car for less than 3 days?

10. A certain brand of light bulb has a lifespan that is normally distributed with a mean (μ) of 1200 hours and a standard deviation (σ) of 150 hours.

DRAFT

(a). What is the probability that a randomly selected light bulb will last less than 1000 hours?

(b). What is the probability that a randomly selected light bulb will last between 1100 hours and 1350 hours?

(c). What is the probability that a randomly selected light bulb will last at most 1250 hours?

(d). If the manufacturer wants to offer a warranty such that only 5% of the light bulbs fail before the warranty expires, what should be the warranty period (in hours)?

DRAFT

Chapter 8

Confidence Interval Estimation

DRAFT

8.1 Introduction

In data science and statistics, estimating population parameters based on sample data is a fundamental task. **Point estimates**, such as the sample mean $\bar{x}$ or sample proportion $\widehat{p}$, provide single-value estimates of these parameters. However, a single point estimate does not convey the uncertainty inherent in the estimation process. This is where interval estimation becomes important.

Interval estimation provides a range of plausible values within which a population parameter is expected to lie. Unlike **point estimation**, which gives only a single value, interval estimation provides an interval that reflects the variability inherent in sampling. This allows researchers to quantify the reliability of their estimates.

In real-world applications, estimating a parameter with a single value is usually insufficient because of natural variability in data. A **confidence interval** provides a range of plausible values for a population parameter and is associated with a specific level of confidence—typically 90%, 95%, or 99%. Confidence intervals are essential in classical statistics because they allow us to assess the precision and reliability of sample-based estimates.

This chapter introduces the concept of confidence interval estimation from the perspective of **classical (frequentist) inference**. We discuss how to construct confidence intervals for key population parameters such as the mean, proportion, variance, and standard deviation. Practical issues, such as determining the required sample size to achieve a desired level of precision, are also addressed. By the end of this chapter, readers will understand how to use con-

fidence intervals to make statistically sound inferences based on sample data.

8.2 Confidence Intervals

A **confidence interval** is a range of values derived from sample data such that, under repeated sampling, a specified proportion of such intervals will contain the true population parameter. It provides a measure of reliability for the sample estimate and conveys the uncertainty due to sampling variability.

The **confidence level**, typically expressed as 90%, 95%, or 99%, represents the long-run proportion of confidence intervals that would include the true population parameter if the sampling process were repeated many times.

- **Confidence Interval (CI):** A range of values, computed from sample data, that contains the true population parameter a specified percentage of the time (e.g., 95%) if the sampling procedure were repeated many times.
- **Confidence Level:** The confidence level (e.g., 95%) represents the proportion of confidence intervals that would include the true parameter if the sampling process were repeated many times.

Illustrative Example

Suppose we want to estimate the average height of adult women in a city. Since measuring every woman is not feasible, we draw a random sample of 100 women. From this sample, we compute the sample mean $\bar{x} = 64$ inches and the sample standard deviation $s = 3$ inches.

Because the estimate is based on a sample, there is uncertainty regarding the true population mean μ. To quantify this uncertainty, we construct a confidence interval. For example, a 95% confidence interval might range from 63.41 to 64.59 inches.

This interval is interpreted in the classical sense: if we repeatedly drew samples of the same size from the population and constructed a 95% confidence interval from each, then approximately 95% of those intervals would contain the true population mean μ.

8.3 Confidence Intervals for the Population Mean

Population Variance Known

When the population variance σ^2 is known, the $(1 - \alpha)$ confidence interval for the population mean μ is:

DRAFT

$$\bar{x} \pm \text{ME} \Rightarrow (\bar{x} - \text{ME}, \bar{x} + \text{ME}) \tag{8.1}$$

where the **margin of error (ME)** is

$$\text{ME} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Here,

- $\bar{x}$ is the sample mean,
- σ is the population standard deviation,
- n is the sample size,
- $z_{\frac{\alpha}{2}}$ is the critical value from the standard normal distribution corresponding to $(1 - \alpha/2)$. The critical value $z_{\frac{\alpha}{2}}$ can be found in Table 8.1.

Margin of Error (ME): The margin of error indicates the maximum expected difference between the sample mean and the true population mean. It determines the width of the confidence interval; a smaller ME corresponds to a more precise estimate.

The confidence interval can also be expressed probabilistically as

$$P\Big(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\Big) = 1 - \alpha,$$

where the probability refers to the long-run proportion of intervals containing the true mean under repeated sampling.

Population Variance Unknown or Small Sample Size

When the population variance σ^2 is unknown and/or the sample size is small ($n < 30$), we use the t-distribution. In this case the confidence interval is

$$\bar{x} \pm \text{ME} \Rightarrow (\bar{x} - \text{ME}, \bar{x} + \text{ME})$$

where,

$$\text{ME} = t_{\frac{\alpha}{2},\nu} \frac{s}{\sqrt{n}}$$

and

- s is the sample standard deviation,
- $\nu = n - 1$ is the degrees of freedom,

DRAFT

- $t_{\frac{\alpha}{2},\nu}$ is the critical value from the t-distribution corresponding to $(1 - \alpha/2)$.

The critical value of t for the desired confidence level can be found in Table 8.3.

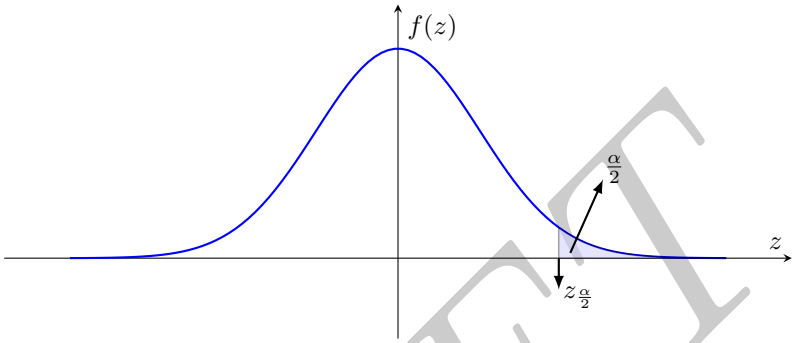


Figure 8.1: Standard Normal Distribution with shaded area to the right of $z_{\frac{\alpha}{2}}$.

Table: The value of $\frac{\alpha}{2}$ with corresponding values $z_{\frac{\alpha}{2}}$.

$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
0.001	3.090	0.021	2.034	0.041	1.739	0.061	1.546	0.081	1.398
0.002	2.878	0.022	2.014	0.042	1.728	0.062	1.538	0.082	1.392
0.003	2.748	0.023	1.995	0.043	1.717	0.063	1.530	0.083	1.385
0.004	2.652	0.024	1.977	0.044	1.706	0.064	1.522	0.084	1.379
0.005	2.576	0.025	1.960	0.045	1.695	0.065	1.514	0.085	1.372
0.006	2.512	0.026	1.943	0.046	1.685	0.066	1.506	0.086	1.366
0.007	2.457	0.027	1.927	0.047	1.675	0.067	1.499	0.087	1.359
0.008	2.409	0.028	1.911	0.048	1.665	0.068	1.491	0.088	1.353
0.009	2.366	0.029	1.896	0.049	1.655	0.069	1.483	0.089	1.347
0.010	2.326	0.030	1.881	0.050	1.645	0.070	1.476	0.090	1.341
0.011	2.290	0.031	1.866	0.051	1.635	0.071	1.468	0.091	1.335
0.012	2.257	0.032	1.852	0.052	1.626	0.072	1.461	0.092	1.329
0.013	2.226	0.033	1.838	0.053	1.616	0.073	1.454	0.093	1.323
0.014	2.197	0.034	1.825	0.054	1.607	0.074	1.447	0.094	1.317
0.015	2.170	0.035	1.812	0.055	1.598	0.075	1.440	0.095	1.311
0.016	2.144	0.036	1.799	0.056	1.589	0.076	1.433	0.096	1.305
0.017	2.120	0.037	1.787	0.057	1.580	0.077	1.426	0.097	1.299
0.018	2.097	0.038	1.774	0.058	1.572	0.078	1.419	0.098	1.293
0.019	2.075	0.039	1.762	0.059	1.563	0.079	1.412	0.099	1.287
0.020	2.054	0.040	1.751	0.060	1.555	0.080	1.405	0.100	1.282

Remark 8.3.1. *When the sample size is greater than or equal to 30 (i.e., $n \geq 30$), the Central Limit Theorem suggests that the sampling distribution of the sample mean is approximately normally distributed, even if the population*

DRAFT

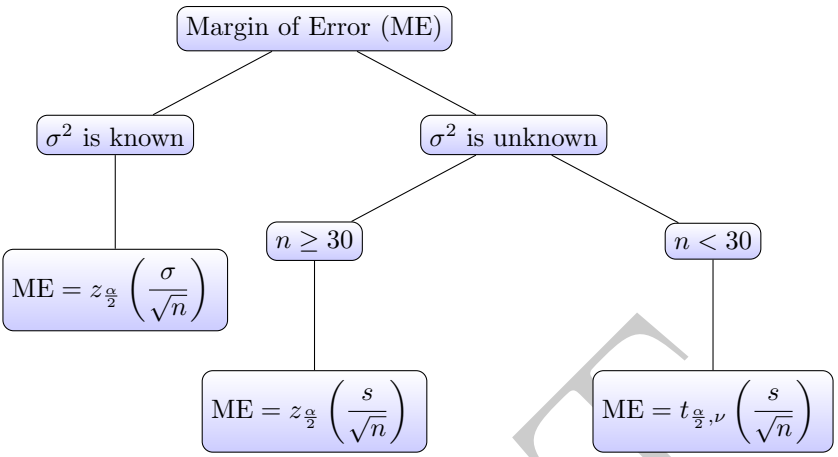


Figure 8.2: The margin of error (ME) for the confidence interval for the population mean.

distribution is unknown. In such cases, it is appropriate to use the critical value $z_{\frac{\alpha}{2}}$ *instead of* $t_{\frac{\alpha}{2},\nu}$.

However, if the sample size is smaller than 30 and the population variance is unknown, it is generally recommended to use the critical value from the t-distribution, denoted as $t_{\frac{\alpha}{2},\nu}$*, which accounts for the additional variability due to the smaller sample size. The flowchart for calculating the margin of error is shown in Figure 8.2.*

Problem 8.1. *A researcher is studying the time spent on social media by teenagers. Based on a previous study, the population variance is known to be* $\sigma^2 = 25$ *minutes*^{2}*. A sample of* $n = 30$ *teenagers shows a sample mean of* $\bar{x} = 120$ *minutes. Construct the following confidence intervals for the population mean and compare their widths and implications:*

(i) 90% confidence interval

(ii) 95% confidence interval

(iii) 99% confidence interval

Solution

Let $x_1, x_2, \ldots, x_n$ denote the times spent on social media by $n = 30$ randomly selected teenagers. These are independent and identically distributed (i.i.d.) random variables with population mean μ and known population variance $\sigma^2 = 25$ minutes^2.

The sample mean is defined as

DRAFT

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We are given that the observed sample mean is $\bar{x} = 120$ minutes.

The confidence interval for the population mean when the population variance is known is

$$\bar{x} \pm \text{ME},$$

where,

$$\text{ME} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

and $z_{\frac{\alpha}{2}}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

(i) 90% Confidence Interval For a 90% confidence level, $z_{\frac{\alpha}{2}} \approx 1.645$. The margin of error is

$$\text{ME} = 1.645 \times \frac{5}{\sqrt{30}} \approx 1.645 \times 0.912 \approx 1.50$$

Thus, the 90% confidence interval is

$$120 \pm 1.50 \Rightarrow (118.50, 121.50)$$

Interpretation: Interpretation: If we repeatedly draw random samples of size 30 and construct a 90% confidence interval each time using this method, then approximately 90% of those intervals will contain the true population mean μ. In practical terms, we say that we are 90% confident that the true average time spent on social media by teenagers lies between 118.50 and 121.50 minutes.

Note: This does not mean that there is a 90% probability that μ lies in this specific interval. In frequentist inference, μ is fixed and the randomness lies in the interval.

(ii) 95% Confidence Interval For a 95% confidence level, $z_{\frac{\alpha}{2}} \approx 1.96$. The margin of error is

$$\text{ME} = 1.96 \times \frac{5}{\sqrt{30}} \approx 1.96 \times 0.912 \approx 1.79$$

Thus, the 95% confidence interval is

$$120 \pm 1.79 \Rightarrow (118.21, 121.79)$$

Interpretation: In repeated sampling, about 95% of such intervals will cover the true population mean μ. In practical terms, based on the observed sample,

DRAFT

we are 90% confident that the true average time spent on social media by teenagers lies between 118.21 and 121.79 minutes.

(iii) 99% Confidence Interval For a 99% confidence level, $z_{\frac{\alpha}{2}} \approx 2.576$. The margin of error is

$$\text{ME} = 2.576 \times \frac{5}{\sqrt{30}} \approx 2.576 \times 0.912 \approx 2.35$$

Thus, the 99% confidence interval is

$$120 \pm 2.35 \Rightarrow (117.65, 122.35)$$

Interpretation: In practice, we say that we are 99% confident that the true average time spent on social media by teenagers lies between 117.65 and 122.35 minutes.

Comparison of Confidence Intervals

- Width of 90% CI: $121.50 - 118.50 = 3.00$ minutes
- Width of 95% CI: $121.79 - 118.21 = 3.58$ minutes
- Width of 99% CI: $122.35 - 117.65 = 4.70$ minutes

Observation: As the confidence level increases, the width of the confidence interval increases. This occurs because a higher confidence level requires a larger $z_{\frac{\alpha}{2}}$-value, which increases the margin of error.

Implications: A narrower interval (lower confidence level) gives a more precise estimate but with lower confidence. A wider interval (higher confidence level) provides higher confidence that the interval contains the true mean, but with less precision. The choice of confidence level depends on the desired balance between precision and certainty.

Problem 8.2. *A factory produces light bulbs with a known population standard deviation of* $\sigma = 100$ *hours. A sample of light bulbs has an average lifespan of* $\bar{x} = 1000$ *hours. Construct a 99% confidence interval for the population mean lifespan of the light bulbs for the following sample sizes:*

(i) $n = 20$

(ii) $n = 30$

(iii) $n = 50$

DRAFT

Solution

Let $x_1, x_2, \ldots, x_n$ denote the lifespans of the randomly selected light bulbs, assumed to be independent and identically distributed (i.i.d.) with population mean μ and known standard deviation $\sigma = 100$ hours. The sample mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The observed sample mean is $\bar{x} = 1000$ hours. For a 99% confidence level, the critical value is

$$z_{\frac{\alpha}{2}} \approx 2.576.$$

(i) Sample size $n = 20$:

$$\text{ME} = 2.576 \cdot \frac{100}{\sqrt{20}} \approx 57.60$$

$$\text{CI} = 1000 \pm 57.60 \Rightarrow (942.40, 1057.60)$$

Interpretation: Under repeated sampling of size $n = 20$, approximately 99% of the resulting intervals would contain the true population mean μ.

(ii) Sample size $n = 30$:

$$\text{ME} = 2.576 \cdot \frac{100}{\sqrt{30}} \approx 47.03$$

$$\text{CI} = 1000 \pm 47.03 \Rightarrow (952.97, 1047.03)$$

Interpretation: With $n = 30$, about 99% of repeated-sample confidence intervals would cover the true mean. Notice the interval is narrower than for $n = 20$, reflecting increased precision.

(iii) Sample size $n = 50$:

$$\text{ME} = 2.576 \cdot \frac{100}{\sqrt{50}} \approx 36.45$$

$$\text{CI} = 1000 \pm 36.45 \Rightarrow (963.55, 1036.45)$$

Interpretation: With a larger sample ($n = 50$), the confidence interval is narrower, providing a more precise estimate of the population mean.

DRAFT

Comparison of Confidence Intervals

- $n = 20$: CI $= (942.40, 1057.60)$, width $= 115.20$
- $n = 30$: CI $= (952.97, 1047.03)$, width $= 94.06$
- $n = 50$: CI $= (963.55, 1036.45)$, width $= 72.90$

Observation: As the sample size increases, the width of the confidence interval decreases. This demonstrates the classical principle that larger samples lead to more precise estimates of the population mean.

Implications: A narrower interval provides a more precise estimate, but all intervals are constructed to maintain 99% confidence. In practice, choosing the sample size depends on the desired precision (width of the interval) and available resources. □

Factors Influencing the Width of a Confidence Interval

The width of a confidence interval for a population parameter depends on several key factors:

(i). **Sample Size (n):** Increasing the sample size reduces the standard error of the sample mean,

$$\text{se}(\bar{x}) = \frac{\sigma}{\sqrt{n}} \text{ (when } \sigma \text{ is known).}$$

A smaller standard error results in a narrower confidence interval, making the estimate more precise. In the classical (frequentist) sense, larger samples produce intervals that, in repeated sampling, more closely "pin down" the true population parameter.

(ii). **Confidence Level** $(1 - \alpha)$**:** A higher confidence level requires a larger critical value $z_{\frac{\alpha}{2}}$ or $t_{\alpha/2,\nu}$, which increases the margin of error:

$$\text{ME} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad t_{\alpha/2,\nu} \frac{s}{\sqrt{n}}.$$

Consequently, higher confidence levels (e.g., 99%) lead to wider intervals, reflecting the classical principle that greater certainty requires a broader interval to maintain the long-run coverage probability.

(iii). **Population Variability (σ^2):** Greater variability in the population increases the standard error. A more dispersed population leads to larger margins of error and therefore wider confidence intervals. Conversely, a population with smaller variance produces more precise (narrower) intervals for the same sample size and confidence level.

DRAFT

Problem 8.3. *Consider a sample of $n = 20$ measurements of blood pressure with a sample mean of $\bar{x} = 130$ mmHg and a sample standard deviation of $s = 15$ mmHg. Construct a 95% confidence interval for the population mean blood pressure.*

Solution

Let $x_1, x_2, \ldots, x_{20}$ denote the blood pressure measurements of the sampled individuals. We assume these measurements are independent and identically distributed (i.i.d.) random variables with population mean μ and unknown standard deviation σ. The sample mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 130 \text{ mmHg}.$$

Since the population standard deviation is unknown and the sample size is small ($n < 30$), we use the Student's t-distribution to construct the confidence interval. The degrees of freedom are

$$\nu = n - 1 = 19.$$

From the t-distribution table, the critical value for a 95% confidence level with 19 degrees of freedom is approximately

$$t_{0.025,19} \approx 2.093.$$

The margin of error (ME) is calculated as

$$\text{ME} = t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} = 2.093 \cdot \frac{15}{\sqrt{20}} \approx 2.093 \cdot 3.354 \approx 7.02$$

Thus, the 95% confidence interval for the population mean is

$$\text{CI} = \bar{x} \pm \text{ME} = 130 \pm 7.02 \Rightarrow (122.98, 137.02)$$

Interpretation: According to classical (frequentist) inference, if we were to take many independent samples of size $n = 20$ from the population and compute a 95% confidence interval for each sample, approximately 95% of those intervals would contain the true population mean blood pressure μ. □

Problem 8.4. *A study measures the daily caffeine intake of $n = 25$ adults. The sample has a mean intake of $\bar{x} = 200$ mg and a sample standard deviation of $s = 50$ mg. Construct a 90% confidence interval for the population mean daily caffeine intake.*

DRAFT

Solution

Let $x_1, x_2, \ldots, x_{25}$ denote the daily caffeine intake of the sampled adults. We assume these measurements are independent and identically distributed (i.i.d.) random variables with population mean μ and unknown standard deviation σ. The sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 200 \text{ mg}.$$

Since the population standard deviation is unknown and the sample size is small ($n < 30$), we use the Student's t-distribution to construct the confidence interval. The degrees of freedom are

$$\nu = n - 1 = 24.$$

From the t-distribution table, the critical value for a 90% confidence level with 24 degrees of freedom is approximately

$$t_{0.05,24} \approx 1.711.$$

The margin of error (ME) is

$$\text{ME} = t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} = 1.711 \cdot \frac{50}{\sqrt{25}} = 1.711 \cdot 10 \approx 17.11$$

Thus, the 90% confidence interval for the population mean is

$$\text{CI} = \bar{x} \pm \text{ME} = 200 \pm 17.11 \Rightarrow (182.89, 217.11)$$

Interpretation: In the frequentist (classical) sense, if we were to take many independent samples of size $n = 25$ and compute a 90% confidence interval for each, approximately 90% of these intervals would contain the true population mean daily caffeine intake μ.

□

Problem 8.5. *A study measures the effectiveness of a new diet on weight loss. A sample of $n = 25$ participants who followed the diet for one month had a sample mean weight loss of $\bar{x} = 4.5$ kg with a sample standard deviation $s = 1.2$ kg. Construct a 90% confidence interval for the population mean weight loss.*

Solution

Let $x_1, x_2, \ldots, x_{25}$ denote the weight loss (in kg) for each participant. We assume these measurements are independent and identically distributed (i.i.d.) with population mean μ and unknown standard deviation σ. The sample mean is

DRAFT

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 4.5 \text{ kg}.$$

Since the population standard deviation is unknown and the sample size is small ($n < 30$), we use the Student's t-distribution to construct the confidence interval. The degrees of freedom are

$$\nu = n - 1 = 24.$$

From the t-distribution table, the critical value for a 90% confidence level with 24 degrees of freedom is approximately

$$t_{0.05,24} \approx 1.711.$$

The margin of error (ME) is calculated as

$$\text{ME} = t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} = 1.711 \cdot \frac{1.2}{\sqrt{25}} = 1.711 \cdot 0.24 \approx 0.411$$

Thus, the 90% confidence interval for the population mean weight loss is

$$\text{CI} = \bar{x} \pm \text{ME} = 4.5 \pm 0.411 \Rightarrow (4.089, 4.911)$$

Interpretation: In the frequentist (classical) sense, if we were to take many independent samples of size $n = 25$ and construct a 90% confidence interval for each, approximately 90% of these intervals would contain the true mean weight loss μ. □

Problem 8.6. *A research team is evaluating the effect of a new vaccine on reducing symptoms of a respiratory illness. In a sample of $n = 30$ vaccinated individuals, the sample mean reduction in symptoms score is $\bar{x} = 6.3$ with a sample standard deviation $s = 1.8$. Construct a 99% confidence interval for the population mean reduction in symptoms score.*

Solution

Let $x_1, x_2, \ldots, x_{30}$ denote the reduction in symptoms score for each vaccinated individual. We assume these measurements are independent and identically distributed (i.i.d.) with population mean μ and unknown standard deviation σ. The sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 6.3.$$

Since the sample size is sufficiently large ($n \geq 30$), we can use the normal distribution to construct the confidence interval. The critical value for a 99% confidence level is

DRAFT

Figure 8.3: A.4: Critical points = $t_{\frac{\alpha}{2},\nu}$ of the t-distribution with its degrees of freedom ($\nu = n - 1$)

ν	$\frac{\alpha}{2}$						
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

$$z_{0.005} \approx 2.576.$$

The margin of error (ME) is calculated as

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 2.576 \cdot \frac{1.8}{\sqrt{30}} = 2.576 \cdot 0.328 \approx 0.845$$

DRAFT

Thus, the 99% confidence interval for the population mean is

$$\text{CI} = \bar{x} \pm \text{ME} = 6.3 \pm 0.845 \Rightarrow (5.455, 7.145)$$

Interpretation: In the classical (frequentist) sense, if we were to take many independent samples of size $n = 30$ from the vaccinated population and construct a 99% confidence interval for each, approximately 99% of these intervals would contain the true mean reduction in symptoms score μ. □

Problem 8.7. *A biologist studies the effect of a new fertilizer on plant growth. A sample of $n = 10$ plants is measured for growth in centimeters over a month. The growth measurements are:*

$$12.5, 15.3, 14.0, 16.7, 13.5, 14.8, 15.1, 13.2, 17.0, 12.0$$

(i). Calculate the sample mean ($\bar{x}$) and sample standard deviation (s) of the growth measurements.

(ii). Using a 95% confidence level, construct a confidence interval for the mean growth of the plants.

Solution

(i) Sample Mean and Standard Deviation: Let $x_1, x_2, \ldots, x_{10}$ denote the growth measurements. Using Table 8.1, the sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{144.1}{10} = 14.41 \text{ cm}.$$

The sample variance is calculated using

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{9}\left(2101.77 - 10 \cdot 14.41^2\right) \approx 2.81$$

and the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{2.81} \approx 1.68 \text{ cm}.$$

DRAFT

Table 8.1: Growth Measurements and Their Squares

i	x_i	x_i^2
1	12.5	156.25
2	15.3	234.09
3	14.0	196.00
4	16.7	278.89
5	13.5	182.25
6	14.8	219.04
7	15.1	228.01
8	13.2	174.24
9	17.0	289.00
10	12.0	144.00
Total	144.1	2101.77

(ii) 95% Confidence Interval: Since $n = 10 < 30$ and the population standard deviation is unknown, we use the t-distribution with $\nu = n - 1 = 9$ degrees of freedom. From the t-table, the critical value for 95% confidence is $t_{0.025,9} \approx 2.262$. The margin of error (ME) is

$$\text{ME} = t_{\frac{\alpha}{2},\nu} \cdot \frac{s}{\sqrt{n}} = 2.262 \cdot \frac{1.68}{\sqrt{10}} \approx 2.262 \cdot 0.531 \approx 1.20$$

Thus, the 95% confidence interval for the population mean growth is

$$\text{CI} = \bar{x} \pm \text{ME} = 14.41 \pm 1.20 \Rightarrow (13.21, 15.61)$$

Interpretation: In the classical sense, if we were to repeat this experiment many times and construct a 95% confidence interval from each sample, approximately 95% of those intervals would contain the true mean growth μ. □

Python Code

```
import numpy as np
import scipy.stats as stats

# Growth measurements in centimeters
growth_measurements = [12.5, 15.3, 14.0, 16.7, 13.5, 14.8,
    15.1, 13.2, 17.0, 12.0]

# (i) Calculate the sample mean and sample standard
    deviation
```

DRAFT

```
sample_mean = np.mean(growth_measurements)
sample_std_dev = np.std(growth_measurements, ddof=1)  # Sample standard deviation

print(f"Sample Mean: {sample_mean:.2f} cm")
print(f"Sample Standard Deviation (s): {sample_std_dev:.2f} cm")

# (ii) Construct a 95% confidence interval for the mean growth
confidence_level = 0.95
n = len(growth_measurements)  # Sample size

# Critical t-value for the given confidence level
t_critical = stats.t.ppf((1 + confidence_level) / 2, df=n - 1)

# Standard error of the mean
standard_error = sample_std_dev / np.sqrt(n)

# Margin of error
margin_of_error = t_critical * standard_error

# Confidence interval
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error

print(f"95% Confidence Interval for the mean growth: ({ci_lower:.2f}, {ci_upper:.2f}) cm")


```

Listing 8.1: The 95% Confidence Interval for the Population Mean.

Problem 8.8. *A study is conducted to determine the average cholesterol level in a population after a new treatment. A sample of $n = 50$ patients shows an average cholesterol level of $\bar{x} = 190$ mg/dL with a sample standard deviation $s = 15$ mg/dL. Construct a 95% confidence interval for the mean cholesterol level.*

Solution: Since $n = 50 > 30$, the sample is large enough to use the normal distribution. For a 95% confidence level, the critical value is

$$z_{\frac{\alpha}{2}} = z_{0.025} \approx 1.96$$

The margin of error (ME) is

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 1.96 \cdot \frac{15}{\sqrt{50}} \approx 1.96 \cdot 2.121 \approx 4.16$$

Hence, the 95% confidence interval for the population mean is

$$\text{CI} = \bar{x} \pm \text{ME} = 190 \pm 4.16 \Rightarrow (185.84, 194.16)$$

Interpretation: If we were to repeat this study many times and construct a 95% confidence interval each time, approximately 95% of those intervals would contain the true population mean cholesterol level μ. □

Problem 8.9. *A clinical trial was conducted to test the effectiveness of a new drug in reducing blood pressure. A sample of $n = 40$ patients who received the drug had an average reduction in blood pressure of $\bar{x} = 8$ mmHg with a sample standard deviation $s = 2.5$ mmHg. Construct a 95% confidence interval for the mean reduction in blood pressure.*

Solution: Since $n = 40 > 30$, we can use the normal distribution. For a 95% confidence level, the critical value is

$$z_{\frac{\alpha}{2}} = z_{0.025} \approx 1.96$$

The margin of error (ME) is

$$\text{ME} = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 1.96 \cdot \frac{2.5}{\sqrt{40}} \approx 1.96 \cdot 0.395 \approx 0.774$$

The 95% confidence interval is

$$\text{CI} = \bar{x} \pm \text{ME} = 8 \pm 0.774 \Rightarrow (7.226, 8.774)$$

Interpretation: If this study were repeated many times, about 95% of the 95% confidence intervals calculated from such samples would contain the true mean reduction in blood pressure μ. □

8.4 Confidence Intervals for Variances and Standard Deviations

In data science, estimating the variance and standard deviation of a population is essential for understanding the variability in data. Confidence intervals provide a range of plausible values for these parameters based on sample data.

8.4.1 Confidence Interval for Variance

To compute a confidence interval for the population variance (σ^2), we utilize the chi-square distribution. The confidence interval can be expressed as

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right). \tag{8.2}$$

DRAFT

That is,

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha,$$

where,

- n = sample size
- s^2 = sample variance
- $\chi^2_{1-\alpha/2}$ = the lower chi-square critical value at $\alpha/2$ with $n-1$ degrees of freedom
- $\chi^2_{\alpha/2}$ = the upper chi-square critical value at $1-\alpha/2$ with $n-1$ degrees of freedom

The critical value of χ^2 can be calculated from the standard chi-square distribution. Some values are given in Table 8.2.

8.4.2 Confidence Interval for Standard Deviation

To derive the confidence interval for the population standard deviation (σ), we simply take the square root of the endpoints of the confidence interval for the variance. Thus the $100(1-\alpha)\%$ confidence interval for σ is

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right). \tag{8.3}$$

That is,

$$P\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right) = 1-\alpha.$$

Important Notes

- These methods assume that the sample comes from a normally distributed population.
- The confidence interval will become wider with a decrease in sample size or an increase in the confidence level.

Problem 8.10. *A sample of* 25 *measurements yields a sample variance of* 20. *Construct a* 95% *confidence interval for the population variance and standard deviation.*

DRAFT

Solution: Here, $s^2 = 20$, $n = 25$, and for a 95% confidence level with 24 degrees of freedom, $\chi^2_{\frac{\alpha}{2}} \approx 39.36$ and $\chi^2_{1-\frac{\alpha}{2}} \approx 12.40$. Using these information, the confidence interval for the variance is

$$\left(\frac{24 \times 20}{39.36}, \frac{24 \times 20}{12.40}\right) = (12.20, 38.71).$$

That is,

$$P\left(12.20 \leq \sigma^2 \leq 38.71\right) \approx 95\%.$$

Thus, the 95% confidence interval for the population variance is $(12.20, 38.71)$ and the 95% confidence interval for the standard deviation is approximately

$$(\sqrt{12.20}, \sqrt{38.71}) = (3.49, 6.22).$$

That is,

$$P\left(3.49 \leq \sigma \leq 6.22\right) \approx 95\%.$$

Problem 8.11. *A biostatistical study is conducted with a sample of* 10 *patients, and the following the systolic blood pressure measurements (in mmHg) are recorded:* $120, 125, 130, 135, 140, 145, 150, 155, 160, 165$. *Calculate the variance and standard deviation of the blood pressure measurements. Also, compute a 95% confidence interval for the variance.*

Solution: First, calculate the mean $\bar{x}$:

$$\bar{x} = \frac{120 + 125 + 130 + 135 + 140 + 145 + 150 + 155 + 160 + 165}{10} = 142.5$$

Next, compute the variance which is

$$s^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right) = \frac{1}{10-1}\left(205125 - 10 \times 142.5^2\right) \approx 229.17$$

i	x_i	x_i^2
1	120	14400
2	125	15625
3	130	16900
4	135	18225
5	140	19600
6	145	21025
7	150	22500
8	155	24025
9	160	25600
10	165	27225
Total	**1425**	**205125**

DRAFT

For the 95% confidence interval for the variance, use the Chi-square distribution. We have

$$\text{CI}_{\sigma^2} = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \right]$$

where $\alpha = 0.05$, $n = 10$, and the degrees of freedom $\nu = n - 1 = 9$.

For degrees of freedom 9, the Chi-square critical values ae

$$\chi^2_{0.025} \approx 19.023 \quad \text{and} \quad \chi^2_{0.975} \approx 2.700$$

$$\text{CI}_{\sigma^2} = \left[\frac{9 \times 229.17}{19.023}, \frac{9 \times 229.17}{2.700} \right] = (108.74, 764.81).$$

That is,

$$P\left(108.74 \leq \sigma^2 \leq 764.81\right) \approx 95\%.$$

The 95% confidence interval for the variance of blood pressure measurements, $[108.74, 764.81]$, indicates that we are 95% confident the true population variance lies within this range.

Python Code

```
import numpy as np
import scipy.stats as stats

# Blood pressure measurements in mmHg
blood_pressure = [120, 125, 130, 135, 140, 145, 150, 155,
   160, 165]

# Calculate variance and standard deviation
variance = np.var(blood_pressure, ddof=1)  # Sample
   variance
standard_deviation = np.std(blood_pressure, ddof=1)  #
   Sample standard deviation

print("Variance: {:.2f} sq. mmHg".format(variance))
print("Standard Deviation: {:.2f} mmHg".format(
   standard_deviation))

# Compute the 95% confidence interval for the variance
n = len(blood_pressure)  # Sample size
alpha = 0.05  # Significance level

# Chi-squared critical values
chi2_lower = stats.chi2.ppf(alpha / 2, df=n - 1)
chi2_upper = stats.chi2.ppf(1 - alpha / 2, df=n - 1)

# Confidence interval for the variance
```

```
ci_lower = (n - 1) * variance / chi2_upper
ci_upper = (n - 1) * variance / chi2_lower

print("95% Confidence Interval for the variance: ({:.2f},
   {:.2f}) sq. mmHg.format(ci_lower, ci_upper))
```

Listing 8.2: The 95% Confidence Interval for the Population Variance.

Problem 8.12. *A clinical trial measures the time (in minutes) taken for a specific treatment to have an effect in* 6 *patients, yielding the following times:* $22, 27, 25, 30, 24, 28$. *Determine the variance and standard deviation of the time data. Also, compute a 99% confidence interval for the variance.*

Solution: First, calculate the mean $\bar{x}$:

$$\bar{x} = \frac{22 + 27 + 25 + 30 + 24 + 28}{6} = 26$$

Next, compute the variance which is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{6-1}\left[(22-26)^2 + (27-26)^2 + (25-26)^2\right. \\ &\qquad \left. +(30-26)^2 + (24-26)^2 + (28-26)^2\right] \\ &= \frac{1}{5}\left[16 + 1 + 1 + 16 + 4 + 4\right] \\ &= \frac{42}{5} = 8.4 \end{aligned}$$

The standard deviation is

$$s = \sqrt{s^2} = \sqrt{8.4} \approx 2.90$$

For the 99% confidence interval for the variance, we have

$$\text{CI}_{\sigma^2} = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right].$$

where $\alpha = 0.01$, $n = 6$, and the degrees of freedom $\nu = n - 1 = 5$. We use the following Chi-square critical values

$$\chi^2_{0.005} \approx 16.75 \quad \text{and} \quad \chi^2_{0.995} \approx 0.41$$

Then, the confidence interval is

$$\text{CI}_{\sigma^2} = \left[\frac{5 \times 8.4}{16.75}, \frac{5 \times 8.4}{0.41}\right] = [2.51, 102.01]$$

DRAFT

That is,

$$P\left(2.51 \leq \sigma^2 \leq 102.01\right) \approx 99\%.$$

The 99% confidence interval for the variance of the treatment effect times, $[2.51, 102.01]$, indicates that we are 99% confident the true population variance lies within this range. This suggests that there is considerable variability in the time taken for the treatment to have an effect among patients.

Python Code

```
import numpy as np
import scipy.stats as stats

# Given times
data = np.array([22, 27, 25, 30, 24, 28])

# Calculate mean
mean = np.mean(data)

# Calculate variance (using ddof=1 for sample variance)
variance = np.var(data, ddof=1)

# Calculate standard deviation
std_dev = np.std(data, ddof=1)

# Degrees of freedom
n = len(data)
df = n - 1

# Chi-square critical values for 99% confidence interval
alpha = 0.01
chi2_lower = stats.chi2.ppf(alpha / 2, df)
chi2_upper = stats.chi2.ppf(1 - alpha / 2, df)

# Confidence interval for variance
ci_lower = (df * variance) / chi2_upper
ci_upper = (df * variance) / chi2_lower

# Output results
print(f"Mean: {mean:.2f}")
print(f"Variance: {variance:.2f}")
print(f"Standard Deviation: {std_dev:.2f}")
print(f"Critical Values: ({chi2_lower:.2f},{chi2_upper:.2f})")
print(f"99% Confidence Interval for Variance: ({ci_lower:.2f}, {ci_upper:.2f})")

```

Listing 8.3: The 99% Confidence Interval for the Population Variance.

8.5 Confidence Intervals for Population Proportions

A confidence interval for a population proportion estimates the true proportion of a population based on sample data. It is calculated using the sample proportion $\widehat{p} = \frac{x}{n}$, where x is the number of successes and n is the sample size.

For a given confidence level (e.g., 95%), the critical value z is determined from the standard normal distribution. The margin of error (ME) is calculated as

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where the quantity

$$\text{se}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is called standard error (se) of $\widehat{p}$. The confidence interval is then constructed as

$$\widehat{p} \pm \text{ME} \Rightarrow (\widehat{p} - \text{ME}, \widehat{p} + \text{ME})$$

That is,

$$P\left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1 - \alpha.$$

This interval provides a range within which we can be confident that the true population proportion lies, based on the sample data. Validity conditions include having a sufficiently large sample size ($n \geq 30$) and ensuring $np \geq 5$ and $n(1-p) \geq 5$.

Confidence intervals are crucial for assessing uncertainty and making informed decisions based on sample estimates.

Problem 8.13. *In a survey of* 500 *voters,* 300 *indicated they would vote for candidate A. Construct a* 95% *confidence interval for the proportion of voters who support candidate A.*

Solution: The sample size is $n = 500$ and the number of voters for candidate A is $x = 300$. The sample proportion is $\widehat{p} = \frac{300}{500} = 0.6$. For a 95% confidence level, the critical value is $z \approx 1.96$. The margin of error (ME) is calculated as

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96\sqrt{\frac{0.6 \times 0.4}{500}} \approx 1.96 \times 0.0219 \approx 0.0429$$

Thus, the 95% confidence interval is

$$\widehat{p} \pm \text{ME} = 0.6 \pm 0.04295 \Rightarrow (0.5571, 0.6429)$$

DRAFT

That is,

$$P\left(0.5571 \leq p \leq 0.6429\right) \approx 95\%.$$

We are 95% confident that the true proportion of voters supporting candidate A is between 55.71% and 64.29%.

□

Problem 8.14. *In a survey of 150 patients, 45 reported that they are satisfied with their current medication. Estimate the proportion of patients satisfied with their medication and calculate a 95% confidence interval for this proportion.*

Solution: Let $\widehat{p}$ be the sample proportion of satisfied patients. The $\widehat{p}$ is

$$\widehat{p} = \frac{45}{150} = 0.30$$

It is given that $n = 150$ and $z \approx 1.96$ for a 95% confidence level. Thus the margin of error is calculated as

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96\sqrt{\frac{0.30 \times 0.7}{150}} = 1.96 \times 0.037 = 0.072$$

Hence the 95% confidence interval for the proportion is

$$0.30 \pm 0.072 \Rightarrow (0.228, 0.372)$$

That is,

$$P\left(0.228 \leq p \leq 0.372\right) \approx 95\%.$$

So, the 95% confidence interval for the population proportion is $(0.228, 0.372)$.

□

Problem 8.15. *A study on the effectiveness of a new drug finds that out of 200 patients, 50 show significant improvement. Estimate the proportion of patients who show improvement and find a 90% confidence interval for this proportion.*

Solution: Let $\widehat{p}$ be the sample proportion of patients showing improvement. The value of $\widehat{p}$ is

$$\widehat{p} = \frac{50}{200} = 0.25$$

It is given that $n = 200$ and $z \approx 1.645$ for a 90% confidence level. Thus, the margin of error is

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.645\sqrt{\frac{0.25 \times 0.75}{200}} = 1.645 \times 0.033 = 0.054$$

Hence the 95% confidence interval for the proportion is

$$0.25 \pm 0.054 \Rightarrow (0.196, 0.304)$$

That is,

$$P\left(0.196 \leq p \leq 0.304\right) \approx 90\%.$$

DRAFT

Python Code

```
import numpy as np
import scipy.stats as stats

# Given data
n = 200  # total number of patients
x = 50   # number of patients showing improvement

# Estimate the proportion
p_hat = x / n

# Confidence level
confidence_level = 0.90
alpha = 1 - confidence_level

# Standard error
se = np.sqrt((p_hat * (1 - p_hat)) / n)

# Z-score for the given confidence level
z_score = stats.norm.ppf(1 - alpha / 2)

# Margin of error
margin_of_error = z_score * se

# Confidence interval
lower_bound = p_hat - margin_of_error
upper_bound = p_hat + margin_of_error

# Output results
print(f"Estimated proportion of patients showing
  improvement: {p_hat:.4f}")
print(f"90% Confidence Interval: [{lower_bound:.4f}, {
  upper_bound:.4f}]")


```

Listing 8.4: The 95% Confidence Interval for the Population Proportion.

□

Problem 8.16. *In a clinical trial, out of 80 participants, 28 reported a side effect from the medication. Estimate the proportion of participants experiencing side effects and compute a 99% confidence interval for this proportion.*

Solution: Let $\widehat{p}$ be the sample proportion of participants experiencing side effects. The value of $\widehat{p}$ is

$$\widehat{p} = \frac{28}{80} = 0.35$$

It is given that $n = 80$ and $z \approx 2.576$ for a 99% confidence level. Thus, the margin of error is

$$\text{ME} = z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 2.576\sqrt{\frac{0.35 \times 0.65}{80}} = 2.576 \times 0.060 = 0.154$$

Hence the confidence interval is

$$0.35 \pm 0.154 \Rightarrow (0.196, 0.504)$$

That is,

$$P\left(0.196 \leq p \leq 0.504\right) \approx 99\%.$$

8.6 Sample Size Estimation

Sample size estimation is crucial for ensuring that studies and experiments in business are designed to provide reliable and valid results. It's a balance between achieving the accuracy needed for reliable conclusions and the practical constraints of resources, such as time and budget. This section covers sample size estimation for estimating a population mean and a population proportion, including business-related examples.

8.6.1 Sample Size for Estimating a Population Mean

When estimating the mean of a population, the required sample size depends on the desired level of confidence, the acceptable margin of error, and the standard deviation of the population. To determine the ideal sample size, we must first consider the following critical components:

- **Confidence Level:** Indicates the degree of certainty that the sample results reflect the true population value. The most common levels are **95% and 99%**. A higher confidence level offers greater assurance but requires a larger sample size.

- **Margin of Error (or Confidence Interval):** Expressed as a plus or minus (±) percentage, this defines the maximum acceptable deviation between the sample results and the true population value. For example, the value ± 3% margin of error implies the result could vary by up to 3 percentage points in either direction. Achieving a smaller, more precise margin of error typically demands a larger sample size.

- **Population Size:** Refers to the total number of individuals in the group under study. For very large populations (usually exceeding 20,000), the impact on sample size becomes negligible, allowing the population to be considered effectively infinite. In contrast, smaller populations necessitate precise size estimates for accurate sample calculations.

DRAFT

Estimating Sample Size with Known Population Variance

To estimate the sample size n required to achieve a desired margin of error ME with a confidence level of $1-\alpha$, use the following formula:

$$n = \left(\frac{z \times \sigma}{\text{ME}}\right)^2 \tag{8.4}$$

where,

- σ is the standard deviation of the population,
- ME is the desired margin of error.
- $z = z_{\frac{\alpha}{2}}$ is the critical value (See the Table **??**) from the standard normal distribution corresponding to a confidence level of $1-\alpha$,
- If σ is unknown and the sample is small, you would typically use the t-distribution instead of the z-score.

This formula allows researchers to determine the **minimum** sample size needed to achieve a specified level of confidence while ensuring that the estimated mean falls within the desired margin of error. It is more precise logically, because in practice, we often **round up** the result to ensure the margin of error doesn't exceed the target. It is particularly useful in cases where prior studies provide reliable estimates of the population standard deviation.

Problem 8.17. *A company wants to estimate the average monthly spending of its customers on their products. The company aims for a margin of error of \$10 with 95% confidence. If the standard deviation of monthly spending is \$50, determine the required sample size.*

Solution: It is given that

$$\sigma = 50, \quad \text{ME} = 10, \quad z = 1.96 \text{ (for 95\% confidence level)}$$

Thus, the sample size is

$$n = \left(\frac{1.96 \times 50}{10}\right)^2 = \left(\frac{98}{10}\right)^2 = (9.8)^2 = 96.04$$

Rounding up, the required sample size is 97. □

Python Code

DRAFT

```python
import scipy.stats as stats
import math

def calculate_sample_size(margin_of_error, standard_deviation, confidence_level):

# Z-score for the given confidence level
z_score = stats.norm.ppf((1 + confidence_level) / 2)

# Calculate required sample size
required_sample_size = (z_score * standard_deviation / margin_of_error) ** 2
return math.ceil(required_sample_size)  # Round up to the nearest whole number

# Example usage
margin_of_error = 10  # Desired margin of error ($)
standard_deviation = 50  # Population standard deviation ($)
confidence_level = 0.95  # Confidence level

required_size = calculate_sample_size(margin_of_error, standard_deviation, confidence_level)
print("Required Sample Size: {}".format(required_size))

```

To use the formula given in Equation (8.4), we need a value for the population standard deviation, σ. Even if σ is unknown, we can still estimate the sample size using a **preliminary** or **planning value** for it. Some practical ways to find this value in the following:

(i) Use the estimated population standard deviation from **previous studies** as the value for σ.

(ii) Run a **pilot study** to gather some initial data. The standard deviation from this sample can then be used as the value for σ.

(iii) Use **judgment** or a "best guess" for the value of σ. For example, estimate the largest and smallest values in the population. The difference between these values gives you a range. A common suggestion is to divide this range by **4** to get a rough estimate of the standard deviation, which can then serve as the value for σ.

Problem 8.18. *A market research firm wants to estimate the average monthly revenue generated by small businesses. A pilot study estimated the standard deviation of revenue to be \$15,000. To ensure a margin of error of \$2,000 with a 90% confidence level, determine the required sample size.*

Solution: It is given that

$$s = 15000, \quad \text{ME} = 2000, \quad z = 1.645 \text{ (for 90\% confidence level)}$$

Using these information, the sample size is

$$\begin{aligned} n &= \left(\frac{1.645 \times 15000}{2000}\right)^2 \\ &= \left(\frac{24675}{2000}\right)^2 \\ &= (12.3375)^2 \\ &= 152.2 \end{aligned}$$

Rounding up, the required sample size is 153.

Problem 8.19. *A company wants to estimate the average amount of time employees spend on training each month. They conduct a pilot study with 15 employees and find that the standard deviation of training time is 8 hours. To achieve a margin of error of 2 hours with a 95% confidence level, determine the required sample size.*

Solution: It is given,

$$s = 8 \text{ hours}, \quad \text{ME} = 2 \text{ hours},$$

For 95% confidence level, the upper critical value $z \approx 1.96$.

$$n = \left(\frac{1.96 \times 8}{2}\right)^2 = \left(\frac{15.68}{2}\right)^2 = (7.84)^2 \approx 61.47$$

Thus, the required sample size is approximately $n = 62$. □

Problem 8.20. *The range for a set of data is estimated to be 36.*

(a). What is the planning value for the population standard deviation?

(b). At 95% confidence, how large a sample would provide a margin of error of 3?

(c). At 95% confidence, how large a sample would provide a margin of error of 2?

Solution: (a.)
The estimated population standard deviation (σ) can be calculated using the range,

$$\sigma \approx \frac{\text{Range}}{4}$$

DRAFT

Given that the range is 36, we have

$$\sigma \approx \frac{36}{4} = 9$$

(b.) The formula for the sample size (n) is

$$n = \left(\frac{z \times \sigma}{\text{ME}}\right)^2$$

where,

- z is the Z-score for 95% confidence ($Z \approx 1.96$)
- σ is the estimated population standard deviation
- ME is the margin of error

Using $\sigma = 9$ and $\text{ME} = 3$, we have

$$n = \left(\frac{1.96 \times 9}{3}\right)^2 = \left(\frac{17.64}{3}\right)^2 = (5.88)^2 \approx 34.57$$

Rounding up, we find

$$n \approx 35$$

(c.) Using the same formula with $\text{ME} = 2$:

$$n = \left(\frac{1.96 \times 9}{2}\right)^2 = \left(\frac{17.64}{2}\right)^2 = (8.82)^2 \approx 77.92$$

Rounding up, we have

$$n = 78$$

□

8.6.2 Sample Size for Estimating a Population Proportion

When designing a study to estimate a population proportion, determining the appropriate sample size is a critical step that balances statistical precision with practical constraints such as time, cost, and resource availability. An inadequately sized sample may produce estimates too imprecise to be useful, while an unnecessarily large sample wastes resources without commensurate gains in precision.

The sample size required to estimate a population proportion with a specified margin of error depends on three key factors: the desired confidence level, the acceptable margin of error, and an estimate of the population proportion

DRAFT

itself. For a confidence level of $(1-\alpha)\times 100\%$, the required sample size is given by,

$$n = \frac{z^2 \times p \times (1-p)}{\text{ME}^2} \tag{8.5}$$

where,

- n is the required sample size
- $z = z_{\frac{\alpha}{2}}$ is the critical value from the standard normal distribution corresponding to the desired confidence level
- p is the estimated population proportion
- ME is the desired margin of error

For commonly used confidence levels, the z-values are approximately 1.96 for 95% confidence, 1.645 for 90% confidence, and 2.576 for 99% confidence. We always round up to ensure the margin of error does not exceed our target.

Problem 8.21. *A retailer wants to estimate the proportion of customers who are satisfied with their new product. The retailer aims for a margin of error of* 0.05 *with 99% confidence. If the estimated proportion of satisfied customers is* 0.6*, determine the required sample size.*

Solution: Given,

$$p = 0.6, \quad \text{ME} = 0.05, \quad z \approx 2.576 \text{ (for 99\% confidence level)}$$

Using these information, the sample sizeis

$$n = \frac{2.576^2 \times 0.6 \times (1-0.6)}{0.05^2} = \frac{6.635 \times 0.6 \times 0.4}{0.0025} = \frac{1.59084}{0.0025} = 636.336$$

Rounding up, the required sample size is 637. □

Problem 8.22. *A health department wants to estimate the vaccination rate in a community. Previous data suggests the rate is around 65%.*

(a) Calculate the sample size needed for a margin of error of 0.03 at 95% confidence.

(b) Calculate the sample size needed for a margin of error of 0.03 at 99% confidence.

(c) How much does the sample size increase when moving from 95% to 99% confidence?

DRAFT

Solution: Given the following information,

- Estimated proportion: $p = 0.65$
- Margin of error: $\text{ME} = 0.03$

(a) For 95% confidence level, $z = 1.96$, we have

$$\begin{aligned} n &= \frac{z^2 \times p \times (1-p)}{\text{ME}^2} \\ &= \frac{(1.96)^2 \times 0.65 \times (1-0.65)}{(0.03)^2} \\ &= \frac{3.8416 \times 0.65 \times 0.35}{0.0009} = \frac{0.8745}{0.0009} = 971.67 \approx 972 \end{aligned}$$

Therefore, a sample size of **972** is needed at 95% confidence level.

(b) For 99% confidence level, $z = 2.576$, we have

$$\begin{aligned} n &= \frac{z^2 \times p \times (1-p)}{\text{ME}^2} \\ &= \frac{(2.576)^2 \times 0.65 \times (1-0.65)}{(0.03)^2} \\ &= \frac{6.6358 \times 0.65 \times 0.35}{0.0009} = \frac{1.5106}{0.0009} = 1678.44 \approx 1679 \end{aligned}$$

Therefore, a sample size of **1679** is needed at 99% confidence level.

(c) The increase in sample size when moving from 95% to 99% confidence:

$$\begin{aligned} \text{Increase} &= 1679 - 972 \\ &= 707 \end{aligned}$$

$$\begin{aligned} \text{Percentage increase} &= \frac{1679 - 972}{972} \times 100\% \\ &= \frac{707}{972} \times 100\% \\ &= 72.74\% \end{aligned}$$

The sample size increases by **707 participants (or approximately 72.74%)** when moving from 95% to 99% confidence. This demonstrates that higher confidence levels require substantially larger sample sizes to maintain the same margin of error. □

DRAFT

When no prior estimate of the population proportion is available, we use the **conservative approach** by setting $p = 0.5$ in the sample size formula. This value of p ensures the calculated sample size is large enough regardless of the true proportion value.

Problem 8.23. *A market researcher wants to estimate the proportion of consumers who prefer a new product with a margin of error of 0.04 and 95% confidence level. No prior information is available about consumer preferences. Calculate the required sample size.*

Solution: Given the following information,

- No prior information: Use $p = 0.5$
- Margin of error: ME $= 0.04$
- Confidence level: 95%, so $z = 1.96$

So, we have

$$n = \frac{z^2 \times p \times (1-p)}{\text{ME}^2} = \frac{(1.96)^2 \times 0.5 \times (1-0.5)}{(0.04)^2} = 600.25 \approx 601$$

Therefore, a sample size of 601 is needed to estimate the population proportion with a margin of error of 0.04 at 95% confidence when no prior information is available.

8.6.3 Sample Size Estimation for Finite Populations

When estimating the mean or proportion of a population, it's important to consider whether the population is **finite** (or small). A general rule of thumb is to consider populations under 20,000 as finite, though this can vary. For a finite population, the standard sample size formulas, which assume an infinite population, can **overestimate** the required sample size. To get a more efficient and accurate sample, we use a correction factor.

To adjust the initial sample size (n_0) for a finite population, we use a finite population correction factor. This correction reduces the sample size needed while maintaining the desired level of precision. When the population size (N) is known, the formula for estimating the final sample size (n) is

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

where,

- n is the final sample size adjusted for the finite population.

DRAFT

- n_0 is the initial sample size calculated without considering the population size. This is usually determined using standard formulas for an infinite population, such as those for a mean or proportion.
- N is the total number of individuals in the population.

Problem 8.24. *A researcher wants to conduct a survey to estimate the proportion of students who use the university's library. The university has a total of 5,000 students. Based on previous studies, the researcher assumes the proportion of students using the library is around 50%. They want to be 95% confident in their results with a margin of error of 5%. First, calculate the initial sample size (n_0) without considering the finite population. Then, adjust this sample size using the finite population correction formula to find the final, corrected sample size (n).*

Solution

Step 1: Calculate the initial sample size (n_0)

The initial sample size for a proportion (for an infinite population) is

$$n_0 = \frac{z^2 \times p \times (1-p)}{\text{ME}^2}$$

where,

- z = Z-score for the desired confidence level. For a 95% confidence level, $z = 1.96$.
- p = estimated proportion. Here, $p = 0.50$.
- ME = 0.05.

Plugging in the values, we have

$$n_0 = \frac{(1.96)^2 \times 0.50 \times (1-0.50)}{(0.05)^2} = \frac{3.8416 \times 0.25}{0.0025} = 384.16$$

So, the initial required sample size is 385 students.

Step 2: Adjust for the finite population (n)

In the second step, we use finite population correction formula to adjust the sample size, given the total population (N) is 5,000.

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

where,

- $n_0 = 385$

DRAFT

- $N = 5000$

Plugging in the values, we have

$$n = \frac{385}{1 + \frac{385}{5000}} = \frac{385}{1 + 0.077} = 357.47$$

Rounding up, the final adjusted sample size (n) is 358.

Note: Instead of surveying 385 students, the researcher only needs to survey 358 students to achieve the desired confidence and margin of error because the population of students is finite. This adjustment results in a more efficient use of resources.

Remark 8.6.1. *When calculating the* ***initial sample size*** n_0*, we usually round to the nearest whole number since we cannot survey a fraction of a person. The method of rounding can depend on the specific context or research guidelines. Both rounding to the nearest whole number and rounding up are acceptable, as long as the reasoning behind the choice is clear.*

8.7 Concluding Remarks

This chapter covered the fundamental concepts of interval estimation, including confidence intervals for means, proportions, and variances. We explored the interpretation and construction of confidence intervals and provided practical examples to illustrate these concepts. Interval estimation is a powerful tool in statistics, offering a range of values for population parameters and providing a measure of reliability for the estimates.

8.8 Chapter Exercises

1. A nutritionist studies the effect of a new diet on weight loss. A sample of 25 participants followed the diet for six weeks, resulting in an average weight loss of 6.2 kg with a known population standard deviation of 1.5 kg. Construct a 95% confidence interval for the mean weight loss of all participants on this diet.

2. An engineer tests the durability of a new type of material. A sample of 15 test pieces showed an average lifespan of 120 hours with a population standard deviation of 20 hours. Calculate the 99% confidence interval for the average lifespan of this material.

3. A teacher measures the effectiveness of a new teaching method. In a sample of 40 students, the average test score was 78 with a known population standard deviation of 10. Construct a 90% confidence interval for the mean test score of all students taught with this method.

DRAFT

4. A pharmaceutical company is evaluating the effectiveness of a new drug. A sample of 50 patients reported an average improvement score of 8.5 on a health scale, with a population standard deviation of 2. Create a 95% confidence interval for the mean improvement score of all patients taking the drug.

5. A sociologist studies the average number of hours people spend on social media. In a sample of 30 individuals, the average time spent was 3.2 hours with a known population standard deviation of 1.1 hours. Construct a 99% confidence interval for the mean time spent on social media by the population.

6. A horticulturist studies the effect of a new watering technique on plant growth. A sample of 10 plants is measured for growth in centimeters over a month. The growth measurements are as follows:

$$11.8, 14.2, 15.0, 13.6, 12.3, 16.1, 14.4, 13.8, 15.7, 12.5$$

   (i). Calculate the sample mean ($\bar{x}$) and sample standard deviation (s) of the growth measurements.

   (ii). Using a 95% confidence level, construct a confidence interval for the mean growth of the plants.

   (iii). If the horticulturist wants to increase the confidence level to 99%, how would that affect the width of the confidence interval? Calculate the new confidence interval.

   (iv). If the growth measurements were to decrease by 1.5 cm for each plant due to an adverse weather condition, how would this change the sample mean and the confidence interval?

   (v). Based on the confidence interval obtained, discuss the effectiveness of the new watering technique. What conclusions can be drawn?

7. A survey of 200 students revealed that 120 of them prefer online classes to in-person classes. Construct a 95% confidence interval for the proportion of all students who prefer online classes.

8. In a clinical trial, 150 patients were given a new treatment, and 90 reported improvement in their condition. Calculate a 99% confidence interval for the proportion of patients who respond positively to the treatment.

9. A marketing firm conducted a survey of 500 customers and found that 275 are satisfied with their services. Using a 90% confidence level, construct a confidence interval for the proportion of all customers who are satisfied.

10. A poll conducted on $1,000$ voters shows that 430 support a certain candidate. Determine the 95% confidence interval for the proportion of all voters who support this candidate.

DRAFT

11. In a study of cholesterol levels in a population, the following measurements (in mg/dL) were recorded: $210, 220, 230, 240, 250, 260$. Calculate the variance and standard deviation of the cholesterol levels. Also, compute a 90% confidence interval for the variance.

12. A researcher wants to estimate the proportion of voters who support a new policy with a margin of error of 0.05 and a confidence level of 95%. If a previous study found that 60% of voters support the policy, calculate the required sample size.

13. A company aims to determine the average time (in minutes) customers spend on their website. They want to estimate the mean with a margin of error of 2 minutes and a confidence level of 99%. If the standard deviation from a pilot study is 10 minutes, find the required sample size.

14. A medical researcher wants to estimate the average blood pressure of patients in a certain age group within 3 mmHg, with a 95% confidence level. From previous studies, the standard deviation is known to be 12 mmHg. Calculate the sample size needed.

15. A social scientist wants to estimate the mean income of households in a city. They want a margin of error of 500 dollars with a confidence level of 90%. If previous research shows that the standard deviation of household income is approximately 3000 dollars, what sample size is necessary?

16. An education researcher plans to conduct a study on student performance and wants to estimate the mean score on a standardized test with a margin of error of 1.5 points. If the standard deviation of test scores is 6 points, calculate the required sample size for a 95% confidence level.

DRAFT

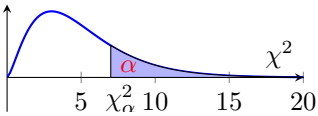


Figure 8.4: Chi-Square Distribution

Table 8.2: A.5: Percentage points of the chi-square distribution ($\chi^2_{\nu,u}$)

ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.07	0.12	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

DRAFT

Chapter 9

Hypothesis Testing for Decision Making

DRAFT

9.1 Introduction

In data science, hypothesis testing is a fundamental aspect of statistical inference, providing a systematic method for making decisions about population parameters based on sample data. This chapter aims to equip readers with a thorough understanding of hypothesis testing, its importance, and its application in various statistical scenarios.

We begin by introducing the basic concepts of hypothesis testing, including the formulation of null and alternative hypotheses and the significance level, which determines the threshold for making decisions. The subsequent sections outline the step-by-step process of conducting hypothesis tests, emphasizing the calculation of test statistics and the determination of acceptance and rejection regions.

A crucial part of this chapter is understanding the role of the p-value in hypothesis testing, as well as its interpretation and significance. We provide illustrative examples involving normal distributions, addressing cases with both known and unknown variances. Additionally, the chapter explores the concept of power in hypothesis testing, demonstrating how to calculate it and why it is vital for ensuring the reliability of test results.

Finally, we cover the methods for estimating the sample size required for mean and proportion tests, highlighting the importance of adequate sample sizes for achieving meaningful and accurate conclusions. The chapter concludes with a set of exercises designed to reinforce the concepts and techniques discussed, providing practical experience in applying hypothesis testing methods.

This comprehensive introduction to hypothesis testing serves as a foundation for more advanced statistical analyses, enabling readers to make informed decisions based on empirical data.

9.2 Concepts of Hypothesis Testing

Consider a scenario where you work for an e-commerce company that has recently redesigned its website to enhance its visual appeal. The marketing team hypothesizes that this new design will lead to an increase in sales. To evaluate this hypothesis, you conduct an A/B test, where half of the visitors are shown the old website (Version A) and the other half are exposed to the new design (Version B).

After running the test for several weeks, you gather data on the number of purchases made by visitors from both versions of the website. Your goal is to determine whether the new design genuinely results in more sales compared to the old version or if any observed differences are attributable to random chance.

Hypothesis testing is a statistical methodology that allows researchers to make inferences about population parameters based on sample data. This method is essential for determining whether observed data can substantiate a specific claim or hypothesis regarding the population. The process begins by formulating two competing hypotheses:

- The null hypothesis, denoted as H_0
- The alternative hypothesis, denoted as H_1 or H_a

The term "null" in "null hypothesis" reflects the concept of a "null effect" or "no effect." It serves as a baseline assumption that there is no significant difference, relationship, or effect between the variables under investigation. Essentially, the null hypothesis is the starting point for statistical testing, positing that any observed differences or effects are due to random chance rather than a true underlying effect. If the evidence is robust enough to reject the null hypothesis, it suggests the presence of a significant effect or relationship that merits further investigation.

Hypothesis: A hypothesis is an assertion or assumption regarding population parameters or characteristics of random variables.

- **Null Hypothesis (H_0):** The null hypothesis asserts that there is no effect or difference between the groups or variables being studied.
- **Alternative Hypothesis (H_1 or H_a):** The alternative hypoth-

DRAFT

esis proposes that there is a significant effect or difference between the groups or variables being studied.

The alternative hypothesis represents the researcher's objective, positing the existence of an effect or difference. For example, a pharmaceutical company may aim to assess whether a new therapy provides greater benefits than an existing treatment. In another case, a health organization might explore differences in physiological indicators between newborns in rural and urban environments. In such studies, researchers formulate hypotheses about the population parameters and utilize sample data to test these hypotheses.

Importance of Hypothesis Testing

Hypothesis testing is crucial because it provides an objective framework for making decisions based on data, ensuring consistency and reproducibility in research. It plays a important role in data analysis for several reasons such as

1. **Avoiding Misinterpretation of Results:** Without hypothesis testing, one might observe that Version B has slightly higher sales than Version A and mistakenly conclude that the new design is definitively superior. Such a conclusion may ignore the possibility that the difference could arise from random variations in the data.

2. **Structured Decision-Making:** Hypothesis testing provides a systematic framework for evaluating whether the difference in sales is statistically significant. It enables a clear approach for determining whether the observed effect is substantial enough to warrant changes to the website design based on the data collected.

3. **Informed Decision-Making:** By employing hypothesis testing, you can make data-driven decisions about implementing the new design site-wide. If the test indicates that the new design significantly increases sales, you may opt to adopt it. Conversely, if the results are not statistically significant, you might decide to retain the old design or explore further improvements.

9.3 Steps for Hypothesis Testing

Hypothesis testing involves the following steps:

Step 1: State the null and alternative hypotheses

Step 2: Set level of significance, usually, denoted by α and critical value

Step 3: Collect Data and define a test statistic and then compute it

DRAFT

Step 4: Construct acceptance ad rejection regions by collecting tabulated value from the sampling distribution fo statistic such as

- standard normal table for Z-test
- standard t-table for T-test
- F-table for F-test
- χ^2-table for χ^2-test, etc.

for the level of significance α or compute p-value and then define decision rule

Step 5: Based on steps 3 and 4, draw a conclusion about H_0.

- reject H_0
  - if the value of test statistic belongs to the rejection region at level of significance $100\alpha\%$.
  - or p-value is less then α.
- otherwise, do not reject H_0 at level of significance $100\alpha\%$.

Details of each step of hypothesis testing for decision-making are explained in the following subsections.

9.3.1 Formulating Hypotheses

A hypothesis becomes a **statistical hypothesis** when it is formulated in a way that allows it to be tested using statistical methods. The first step in hypothesis testing is formulating statistical hypotheses, which consist of the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a). Formulating the null and alternative hypotheses involves clearly defining the research question and the population parameter of interest. The hypotheses must be mutually exclusive and collectively exhaustive, meaning that one of them must be true, and both cannot be true simultaneously.

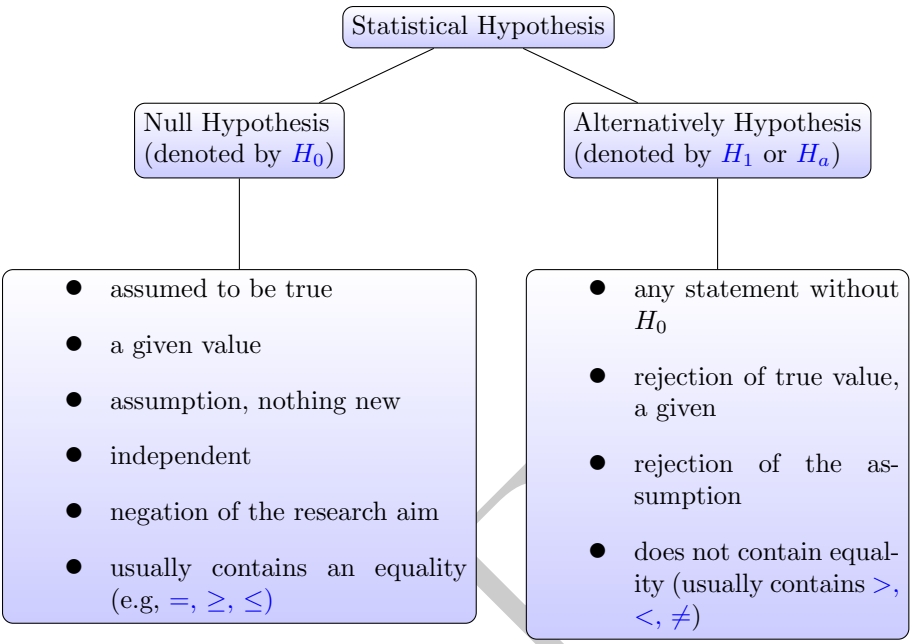
Null Hypothesis (H_0)

The null hypothesis represents a statement of no effect, no difference, or the status quo. It is an assertion that any observed difference or effect in the data is due to random chance rather than a real underlying effect. For example, if a pharmaceutical company is testing a new drug, the null hypothesis might state that the new drug has no effect on patients compared to the existing treatment. Mathematically, the null hypothesis often includes an equality sign, such as

$$H_0 : \theta = \theta_0$$

where θ is the population parameter and θ_0 is the hypothesized value of population parameter.

DRAFT



Alternative Hypothesis (H_1 or H_a)

The alternative hypothesis represents a statement that contradicts the null hypothesis. It indicates the presence of an effect, a difference, or a relationship that the researcher aims to detect. The alternative hypothesis is what the researcher hopes to support through evidence from the data. It can be one-sided or two-sided, depending on the research question. For example, if the pharmaceutical company believes the new drug is more effective, the alternative hypothesis might state that the new drug has a greater effect than the existing treatment

$$H_1 : \theta > \theta_0$$

In two-sided tests, the hypothesis might state that there is a difference in either direction

$$H_1 : \theta \neq \theta_0$$

9.3.2 Level of Significance

The second step in hypothesis testing involves establishing the **level of significance**, denoted by α. This threshold is chosen before conducting the test and represents the probability of rejecting the null hypothesis (H_0) when it is actually true, known as a **Type I error**. Common values for α include 0.05, 0.01, and 0.10, with 0.05 being the most prevalent. For example, if α is set at 0.05, there is a 5% chance of incorrectly concluding that a significant effect exists when it does not. This chosen α level reflects the researcher's tolerance

DRAFT

for making an erroneous decision. A lower α means a stricter criterion for rejecting H_0, which reduces the risk of a Type I error but may increase the risk of a Type II error—failing to reject a false null hypothesis. This threshold thus sets the stage for evaluating the test results, ensuring that decisions align with the predetermined risk level.

Level of Significance: The level of significance is the threshold probability of rejecting the null hypothesis when it is true.

Type I and Type II Errors

Understanding Type I and Type II errors is crucial for interpreting hypothesis testing results. A Type I error occurs when the null hypothesis is rejected when it is true, with the probability of this error being denoted by α. For instance, if $\alpha = 0.05$, the chance of incorrectly rejecting H_0 is 5%. Conversely, a Type II error occurs when the null hypothesis is not rejected when it is false, with the probability of this error denoted by β. The complement of β (i.e., $1 - \beta$) represents the power of the test, which measures the probability of correctly rejecting a false null hypothesis. See the Table 9.1. The trade-off between α and β is critical; reducing α to decrease Type I errors may increase β, making it more likely to miss a true effect. Therefore, choosing appropriate values for α and β involves balancing the risks of both types of errors in the context of the study.

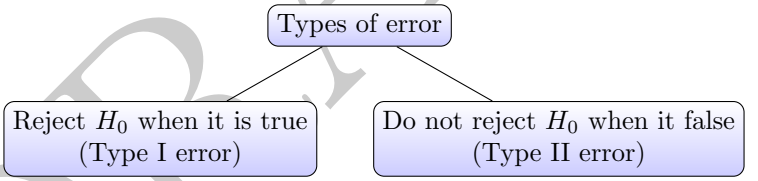


Table 9.1: Level of Significance and Power of the test.

		In Reality	
		H_0 **is TRUE**	H_0 **is FALSE**
Decision	**Accept** H_0	**Correct Decision** $1 - \alpha$ = Confidence level	Type II error $\beta = P$(Type II Error)
	Reject H_0	Type I Error $\alpha = P$(Type I Error)	**Correct Decision** $1 - \beta$ =Power of the test

DRAFT

9.3.3 Test Statistics

The test statistic provides a standardized value that is calculated from sample data during a hypothesis test. It quantifies the discrepancy between the observed sample data and the null hypothesis. The test statistic is a **random variable** because it is derived from random sample data, and it follows a probability distribution. In general, the test statistic for the location parameter test defined in null hypothesis $H_0 : \theta = \theta_0$ is

$$\text{Test statistic} = \frac{\widehat{\theta} - \theta_0}{\text{se}(\widehat{\theta})}$$

where θ is a location parameter. It could be a mean, median, quantile, etc.

9.3.4 Acceptance and Rejection Regions

In hypothesis testing, we define acceptance and rejection regions to decide whether to reject the null hypothesis H_0. The acceptance region is the range of values for which we fail to reject the null hypothesis H_0. The rejection region is the range of values for which we reject H_0. The critical value z_{cri} defines the boundary between these regions. Figure 9.1, Figure 9.2, and Figure 9.3 show the acceptance and rejection regions for a left-tailed test, a right-tailed test, and a two-tailed test, respectively, at a significance level of α.

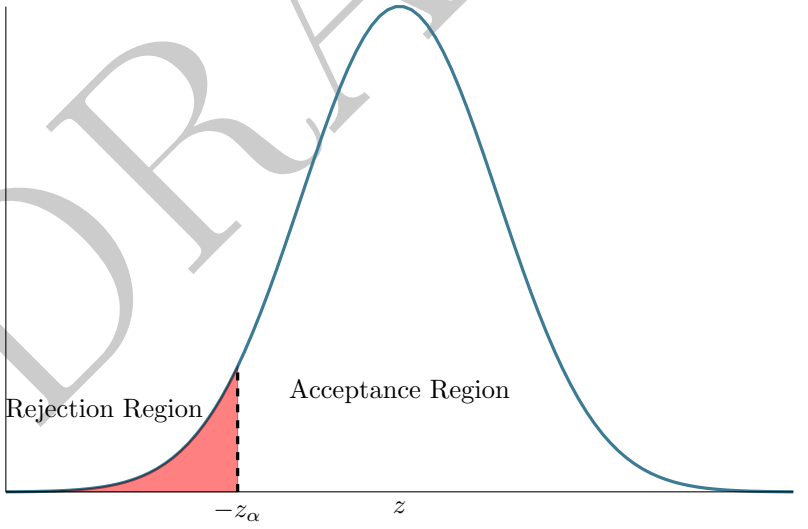


Figure 9.1: Acceptance and Rejection Regions for a Left-Sided Test at $\alpha = 0.05$.

DRAFT

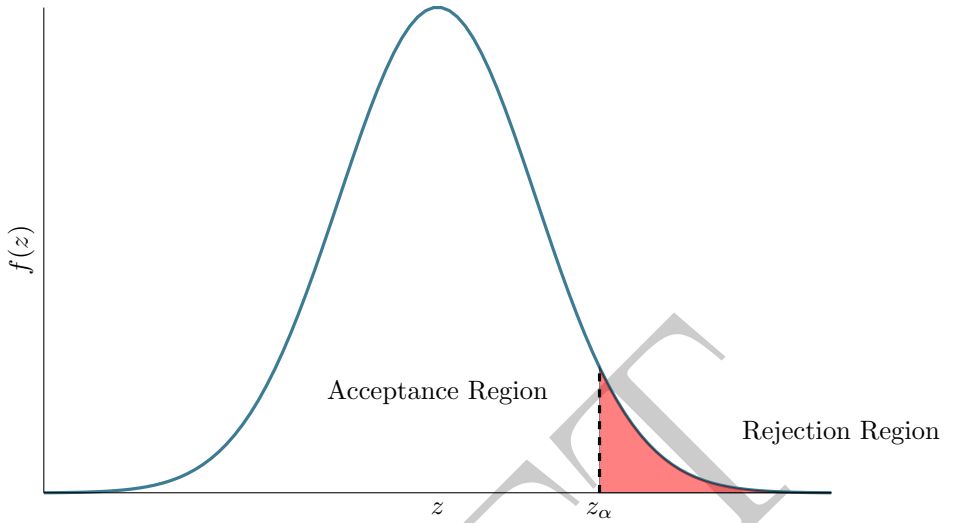


Figure 9.2: Acceptance and Rejection Regions for a Right-Tailed Test at $\alpha = 0.05$.

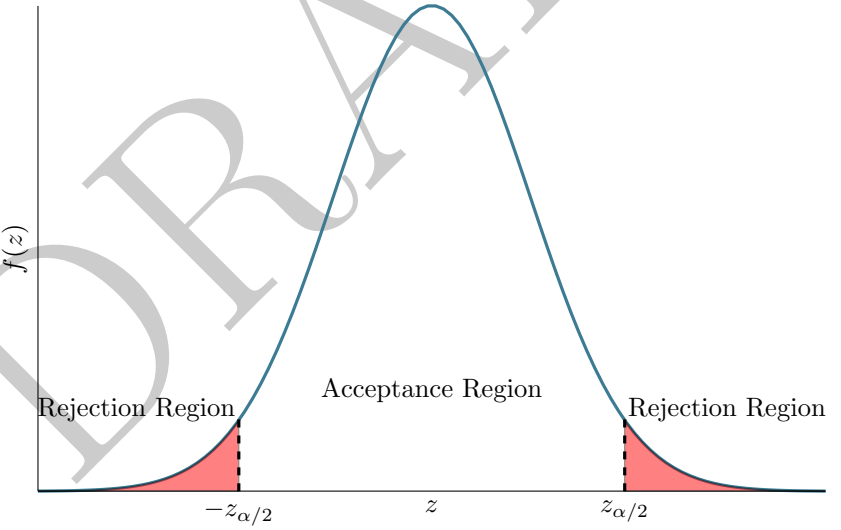


Figure 9.3: Acceptance and Rejection Regions for a Two-Sided Test at $\alpha = 0.05$.

The graphs above represent the standard normal distribution (for Z-test), which is a common distribution used in hypothesis testing. The curve shown is the probability density function of the standard normal distribution, $N(0, 1)$, which is symmetric about the mean $\mu = 0$.

DRAFT

- **Critical Value ($z_{\mathbf{cri}}$)**: Critical values help determine whether to reject the null hypothesis H_0 based on the significance level α and test type.
  - **One-Tailed Test**: For α, the critical value is z_α (upper tail) or $-z_\alpha$ (lower tail). E.g., for $\alpha = 0.05$, $z_\alpha = 1.645$. Reject H_0 if the test statistic exceeds the critical value.
  - **Two-Tailed Test**: For α, the critical values are $\pm z_{\alpha/2}$. E.g., for $\alpha = 0.05$, $z_{\alpha/2} \approx 1.96$. Reject H_0 if the test statistic is outside $\pm z_{\alpha/2}$.
- **Acceptance Region**: Area to the left of z_α where H_0 is not rejected.
- **Rejection Region**: Area to the right of z_α where H_0 is rejected.

Interpreting the Regions

- If the test statistic is in the acceptance region, do not reject H_0 (the data is consistent with H_0).
- If the test statistic is in the rejection region, reject H_0 (the data is unlikely under H_0).

9.3.5 Decision Rules

In hypothesis testing, there are two common methods for making decisions: the critical-value method and the p-value method. The critical-value method involves comparing the test statistic to a critical value. The decision rules are:

Critical-Value Method:

- **One-tailed test (right-tailed)**: Reject H_0 if the test statistic is greater than the critical value; otherwise, fail to reject H_0.
- **One-tailed test (left-tailed)**: Reject H_0 if the test statistic is less than the critical value; otherwise, fail to reject H_0.
- **Two-tailed test**: Reject H_0 if the test statistic is outside the critical values; otherwise, fail to reject H_0.

The p-value Method:

- Reject H_0 if the p-value is less than or equal to α; otherwise, fail to reject H_0.

DRAFT

9.4 The p-value

In hypothesis testing, the p-value helps us determine the strength of the evidence against the null hypothesis, H_0. It is represented by the shadow are in Figure 9.4 in hypothesis testing. The p-value is the area under the curve to the right of the test statistic in a right-tailed test (as shown in the graph). It quantifies the evidence against the null hypothesis: a smaller p-value indicates stronger evidence against H_0.

The p-value: The p-value represents the probability of observing the test statistic as extreme as, or more extreme than, the value observed if the null hypothesis is true. A small p-value (less than α) indicates that the observed data is unlikely under the null hypothesis, leading us to reject H_0.

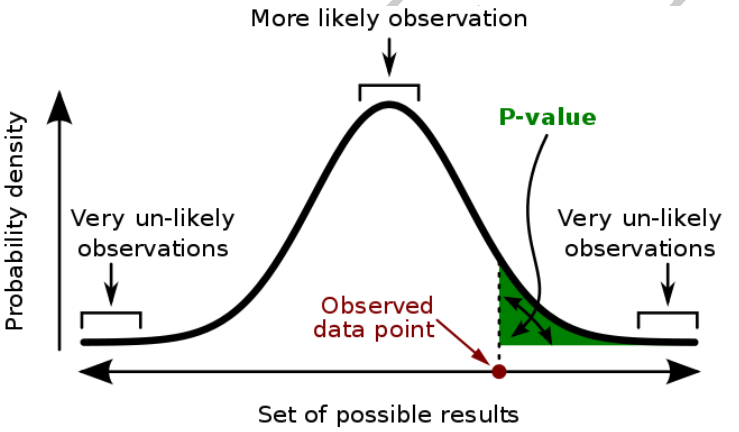


Figure 9.4: The p-value is showed in shadow area.

Interpreting the p-value

- If the p-value is less than or equal to the significance level α, we reject the null hypothesis H_0. This indicates that the observed data is unlikely under H_0.
- If the p-value is greater than α, we fail to reject the null hypothesis H_0. This means there is insufficient evidence to conclude that the observed data is inconsistent with H_0.

Guidelines for Judging the Significance of a p-Value

- If $0.01 \leq p\text{-value} < 0.05$, then the results are significant.

- If $0.001 \leq p$-value < 0.01, then the results are highly significant.
- If p-value < 0.001, then the results are very highly significant.
- If p-value > 0.05, then the results are considered not statistically significant (sometimes denoted by NS).

However, if $0.05 < p$-value < 0.10, then a trend toward statistical significance is sometimes noted.

9.5 Hypothesis Testing for Means

This section focuses on hypothesis tests specifically for means, which are essential for making inferences about the average values of populations or comparing means across different groups. We will explore three key types of tests:

- **One-Sample Test of Means:** This test assesses whether the mean of a single sample differs from a known or hypothesized population mean. It is useful when we want to determine if the sample mean is consistent with a particular value.
- **Testing Equality of Two Means:** Here, we compare the means of two independent samples to evaluate whether there is a significant difference between them. This is crucial when examining the effects of different treatments or conditions on two separate groups.
- **Testing Equality of Several Means:** This test extends the comparison to more than two groups, determining whether there are significant differences among the means of multiple samples. It is commonly used in experiments with multiple treatment levels or groups.

Each of these tests plays a critical role in statistical inference, helping researchers and analysts make data-driven decisions and understand underlying patterns. We will delve into the methodology, assumptions, and applications of these tests in the following subsections.

DRAFT

9.5.1 One-Sample Test of Means

A one-sample test of means is a statistical method used to determine whether the average (mean) of a single sample is significantly different from a known population mean. This test is particularly useful for assessing whether sample data aligns with a specific standard or expectation.

Formulating Hypotheses

Before conducting the test, researchers establish two hypotheses:

- **Null Hypothesis (H_0)**: The sample mean is equal to the hypothesized population mean.
$$H_0 : \mu = \mu_0$$
- **Alternative Hypothesis (H_a)**: The sample mean is not equal to the hypothesized population mean.
$$H_a : \mu \neq \mu_0$$
This is a two-tailed test because we are interested in deviations in both directions (greater or less).

Assumptions

Before performing the test, several key assumptions must be considered:

1. **Normality**: The sample data should be approximately normally distributed, particularly if the sample size is small (less than 30).
2. **Independence**: Each observation within the sample must be independent of others.
3. **Sample Size**: Larger samples enhance the approximation of a normal distribution.

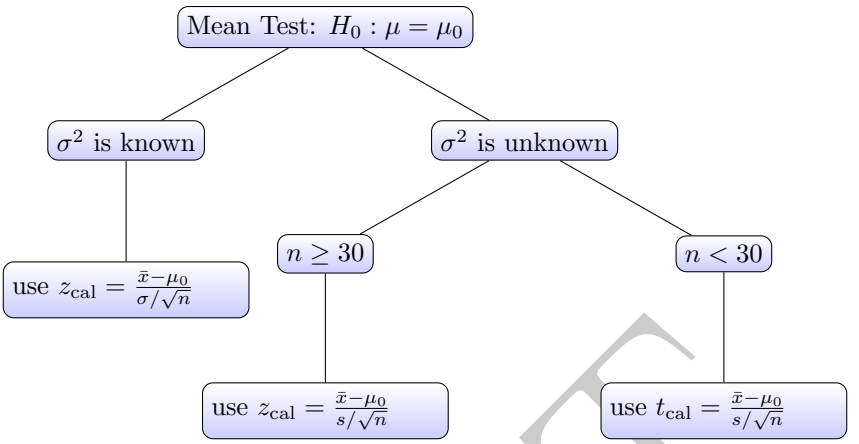
Choosing the Test

Researchers have two primary options for the statistical test:

- **T-test**: This test is appropriate for small samples (usually, $n < 30$) when the population variance is unknown. For T-test, the statistic is
$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
where
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
is the sample standard deviation.
- **Z-test**: This test is suitable for larger samples (usually, $n \geq 30$) or when the population variance is known. For Z-test, the test statistic is
$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In these formulas, $\bar{x}$ represents the sample mean, μ_0 is the population mean being compared against, s is the sample standard deviation (for the T-test), and σ is the population standard deviation (for the Z-test).

DRAFT



Critical-Value Method

Next, researchers determine a critical value using statistical tables based on the chosen significance level (commonly set at 0.05). This value aids in deciding whether to reject the null hypothesis. Alternatively, the p-value is calculated to indicate the probability of observing the results if the null hypothesis is true.

Decision Rule for Z-test

Under H_0, the sampling distribution of Z follows a standard normal distribution. Therefore, for a two-sided alternative hypothesis, we select the critical value $z_{\frac{\alpha}{2}}$ from the standard normal distribution. For a one-tailed test with significance level α, the critical value is denoted as z_α (for the upper tail) or $-z_\alpha$ (for the lower tail). For example:

$$z_\alpha = 1.645 \text{ for } \alpha = 0.05.$$

Suppose the realized value of the test statistic (Z) is z_{cal}. Then,

- for a two-sided test, reject H_0 if and only if $z_{\text{cal}} \geq z_{\frac{\alpha}{2}}$ or $z_{\text{cal}} \leq -z_{\frac{\alpha}{2}}$.
- for a one-sided test, reject H_0 if and only if
  - $z_{\text{cal}} > z_\alpha$ for an upper-tail test, or
  - $z_{\text{cal}} < -z_\alpha$ for a lower-tail test.
  - Equivalently, $|z_{\text{cal}}| > z_\alpha$.
- If the p-value is less than or equal to the significance level (α), reject the null hypothesis H_0.

DRAFT

Decision Rule for T-test

Under H_0, the sampling distribution of the test statistic T follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom, where n is the sample size. Therefore, for a two-sided alternative hypothesis, we select the critical value $t_{\frac{\alpha}{2},(n-1)}$ from the Student's t-distribution with the appropriate degrees of freedom $(n-1)$. For a one-sided alternative hypothesis, the critical value is denoted by $t_{\alpha,(n-1)}$. Suppose the realized value of the test statistic (T) is t_{cal}. Then,

- for a two-sided test, reject H_0 if and only if $t_{\text{cal}} \leq -t_{\frac{\alpha}{2},(n-1)}$ or $t_{\text{cal}} \geq t_{\frac{\alpha}{2},(n-1)}$.
- For a one-sided test, reject H_0 if and only if
  - $t_{\text{cal}} > t_{\alpha,(n-1)}$ for an upper-tail test, or
  - $t_{\text{cal}} < -t_{\alpha,(n-1)}$ for a lower-tail test.
  - Equivalently, $|t_{\text{cal}}| > t_{\alpha,(n-1)}$.

The p-Value Method: Decision Rule

The p-value is calculated to indicate the probability of observing the results if the null hypothesis is true. If the p-value is less than or equal to the significance level (α), reject the null hypothesis H_0. To compute the p-value for a T-test, we can use the following formula based on the t-distribution.

- For a two-tailed test:

$$p\text{-value} = P(T \leq -t_{\text{cal}}) + P(T \geq t_{\text{cal}}) = 2 \times P(T > |t_{\text{cal}}|)$$

- For a one-tailed test (greater than):

$$p\text{-value} = P(T > t_{\text{cal}})$$

- For a one-tailed test (less than):

$$p\text{-value} = P(T < t_{\text{cal}})$$

where T follows a t-distribution with degrees of freedom (df):

$$df = n - 1$$

```
from scipy import stats

# Calculate p-value for a two-tailed test
p_value_two_tailed = 2 * (1 - stats.t.cdf(abs(t_cal), df))
print("Two-Tailed p-value:", p_value_two_tailed)

```

DRAFT

```
# Calculate p-value for a one-tailed test (greater than)
p_value_one_tailed_greater = 1 - stats.t.cdf(t_cal, df)
print("One-Tailed p-value (greater than):",
  p_value_one_tailed_greater)

# Calculate p-value for a one-tailed test (less than)
p_value_one_tailed_less = stats.t.cdf(t_cal, df)
print("One-Tailed p-value (less than):",
  p_value_one_tailed_less)
```

To compute the p-value for Z-test, we can use the following formulas:

- For a two-tailed test:

$$p\text{-value} = 2 \times P(Z > |z_{\text{cal}}|)$$

- For a one-tailed test (greater than):

$$p\text{-value} = P(Z > z_{\text{cal}})$$

- For a one-tailed test (less than):

$$p\text{-value} = P(Z < z_{\text{cal}})$$

Where Z follows a standard normal distribution.

```
from scipy import stats

# Calculate p-value for a two-tailed test
p_value_two_tailed = 2 * (1 - stats.norm.cdf(abs(z_cal)))
print("Two-Tailed p-value:", p_value_two_tailed)

# Calculate p-value for a one-tailed test (greater than)
p_value_one_tailed_greater = 1 - stats.norm.cdf(z_cal)
print("One-Tailed p-value (greater than):",
  p_value_one_tailed_greater)

# Calculate p-value for a one-tailed test (less than)
p_value_one_tailed_less = stats.norm.cdf(z_cal)
print("One-Tailed p-value (less than):",
  p_value_one_tailed_less)
```

Remark 9.5.1. *If σ^2 is known, we always use the Z-test. If σ^2 is unknown, then we have to check if the sample size n is large or not. In this case, we replace the population standard deviation with the sample standard deviation s, and:*

(i). if $n < 30$, we use the T-test,

(ii). if $n \geq 30$, we can use the Z-test.

AFT

Problem 9.1. *From long-term experience, a factory owner knows that the average time required to produce a product is 89 minutes. Factory owner suspects that on Sunday mornings the production time is longer than usual. To investigate this suspicion, a random sample of 12 independent production times is recorded specifically on Sunday mornings. The sample mean is found to be* $\bar{x} = 92.2$ *minutes. Assume that the population variance of production time is known to be* $\sigma^2 = 144$ *minutes*^{2}*. At the 5% level of significance, test whether the mean production time on Sunday mornings is significantly greater than 89 minutes. State your conclusion in the context of the problem.*

Solution

To test whether the production time on Sunday morning is significantly longer, a sample of size $n = 12$ was taken, yielding a sample mean $\bar{x} = 92.2$. It is assumed that the production time is normally distributed with a known variance $\sigma^2 = 144$.

We start by setting up the hypothesis test:

$$H_0 : \mu = 89 \quad \text{vs.} \quad H_1 : \mu > 89.$$

Next, we select the significance level $\alpha = 0.05$. The test statistic is calculated as follows:

$$z_{\text{cal}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{92.2 - 89}{\sqrt{144/12}} = 0.9237.$$

For a right-tailed test at significance level $\alpha = 0.05$, the critical value from the standard normal distribution is $z_\alpha = z_{0.05} = 1.645$.

The rejection region is presented in Figure 13.2. Therefore the region rule is reject H_0 if $z_{\text{cal}} \geq 1.645$.

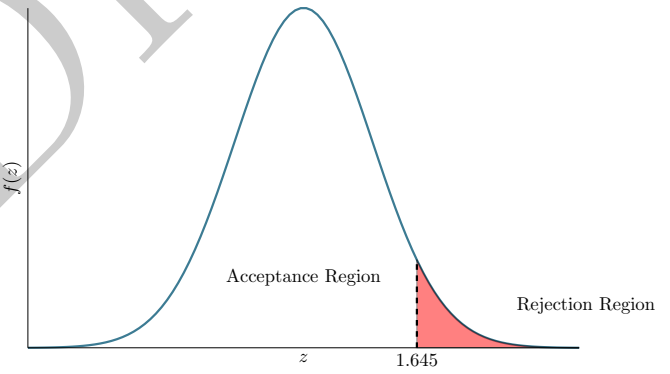


Figure 9.5: Acceptance and Rejection Regions for a Right-Tailed Test at $\alpha = 0.05$.

DRAFT

Since $z_{\text{cal}} = 0.9237$ is less than 1.645, we do not reject the null hypothesis $H_0 : \mu = 89$ at the 5% significance level. Thus, there is insufficient evidence to conclude that the average time required to produce a product is longer than 89 minutes on Sunday mornings.

Alternatively, we can also use the p-value method. The p-value is calculated as:

$$p\text{-value} = P(Z \geq z_{\text{cal}}) = P(Z \geq 0.9237) = 1 - \Phi(0.9237) = 0.1778.$$

This means that if the null hypothesis is true, there's about a 17.78% chance of observing a sample mean of 92.2 minutes or higher due to random sampling variation alone. Since the p-value is greater than 0.05, we do not reject the null hypothesis H_0 at the 5% significance level. This confirms that there is not enough evidence to suggest that production time is longer on Sunday morning.

Problem 9.2. *(**Cardiovascular Disease**) We want to compare fasting serum-cholesterol levels between recent Asian immigrants and the general U.S. population. Cholesterol levels for U.S. women aged 21-40 are normally distributed with a mean of 190 mg/dL. We are interested in whether the mean cholesterol level for recent female Asian immigrants is different from 190 mg/dL. Blood tests from 200 female Asian immigrants show a mean of 181.52 mg/dL and a standard deviation of 40. What conclusions can we draw from this data?*

Solution

To test whether the mean cholesterol level differs from 190, we set up the following hypotheses:

$$H_0 : \mu = 190 \quad \text{vs.} \quad H_1 : \mu \neq 190.$$

Assuming that X follows a normal distribution, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, with a sample mean $\bar{x} = 181.52$, sample standard deviation $s = 40$, and sample size $n = 200$, we proceed with the following steps.

First, we set the significance level to $\alpha = 0.05$. The test statistic is calculated as:

$$z_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{181.52 - 190}{40/\sqrt{200}} = -2.998.$$

For a significance level of $\alpha = 0.05$, the critical value from the z-distribution is $z_{\alpha/2} = z_{0.025} = 1.96$. The rejection region is presented in Figure 9.6. Therefore, the rejection rule is to reject H_0 if $z_{\text{cal}} \leq -1.96$ or $z_{\text{cal}} \geq 1.96$, which is equivalent to rejecting H_0 if $|z_{\text{cal}}| \geq 1.96$.

DRAFT

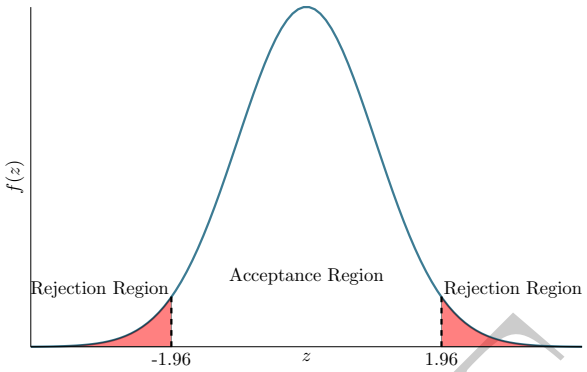


Figure 9.6: Acceptance and Rejection Regions for a Two-Sided Test at $\alpha = 0.05$.

Since $|z_{\text{cal}}| > 1.96$, we reject the null hypothesis $H_0 : \mu = 190$ at the 5% significance level. Thus, we conclude that the mean cholesterol level of recent Asian immigrants is significantly different from that of the general U.S. population.

Alternatively, we compute the p-value for the test statistic. The p-value is given by:

$$p\text{-value} = 2P(Z \geq |z_{\text{cal}}| \mid H_0) = 2P(Z \geq |-2.998|) = 2 \times 0.00137 = 0.0027$$

Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis $H_0 : \mu = 190$ at the 5% level of significance. Thus, we conclude that the mean cholesterol level of recent Asian immigrants is significantly different from that of the general U.S. population.

Problem 9.3. ***(Obstetrics Problem)*** *We want to test if mothers with low socioeconomic status (SES) have babies with lower birthweights than the national average. From 100 consecutive full-term births in a low-SES area, the average birthweight is 115 oz with a standard deviation of 24 oz. Nationally, the mean birthweight is 120 oz. Can we conclude that the mean birthweight in this hospital is lower than the national average?*

Solution

To determine whether the average birthweight is significantly less than 120, we set up the following hypotheses:

$$H_0 : \mu = 120 \quad \text{vs.} \quad H_1 : \mu < 120.$$

Assuming that X follows a normal distribution, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, with a sample mean $\bar{x} = 115$, sample standard deviation $s = 24$, and sample size $n = 100$, we proceed as follows.

First, we set the significance level to $\alpha = 0.05$.

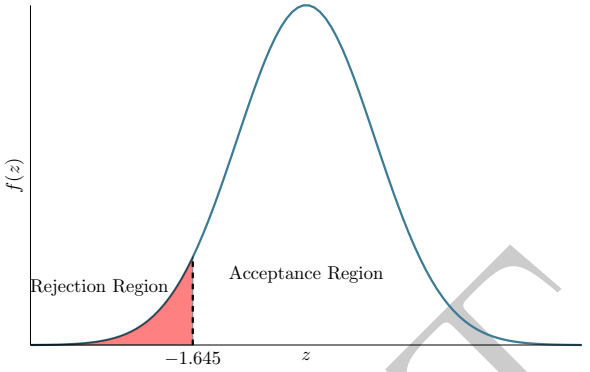


Figure 9.7: Acceptance and Rejection Regions for a Left-Sided Test at $\alpha = 0.05$.

We set the significance level to $\alpha = 0.05$. Since the sample size in this case is large ($n = 100$), we could use either a T-test or a Z-test. However, we typically use the Z-test in this situation. Therefore, the realized value of test statistic is calculated as:

$$z_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{115 - 120}{24/\sqrt{100}} = -2.08.$$

For a one-tailed test at $\alpha = 0.05$, the left-sided critical value is $z_{\text{cri}} = -1.645$. The critical region is presented in Figure 9.7. Thus, the rejection rule is to reject H_0 if $z_{\text{cal}} \leq -1.645$, which is equivalent to rejecting H_0 if $|z_{\text{cal}}| \geq 1.645$.

Since $|z_{\text{cal}}| = 2.08$ is greater than 1.645, we reject the null hypothesis $H_0 : \mu = 120$ at the 5% level of significance. Therefore, we conclude that the average birthweights are lower than the normal average.

Problem 9.4. *From long-term experience, a factory owner knows that the mean time required to produce a product is 89 minutes. He suspects that on Sunday mornings the production time is longer than usual. To investigate this suspicion, a random sample of 12 independent production times is recorded specifically on Sunday mornings. The sample mean is found to be $\bar{x} = 92.2$ minutes, and the sample standard deviation is $s = 10.75$ minutes. At the 5% level of significance, test whether the mean production time on Sunday mornings is significantly greater than 89 minutes. State your conclusion in the context of the problem.*

Solution

We assume that the production time follows a normal distribution. To verify whether the impression that it takes longer on Monday morning is correct, we

DRAFT

conduct a hypothesis test at a significance level of 5%.

We test the following hypotheses:

$$H_0 : \mu = 89$$

$$H_1 : \mu > 89$$

Assuming X follows a normal distribution with an unknown variance σ^2, the test statistic is given by

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)} \text{ under } H_0.$$

We use the significance level α to find the critical value $t_{\alpha,(n-1)}$. From the Student's t-distribution table (Table A6 in the Appendix), we find that $t_{\alpha,(n-1)} = t_{0.05,11} = 1.795$.

The decision rule is to reject H_0 if $|t_{\text{cal}}| > t_{\alpha,(n-1)} = 1.795$. Given that $n = 12$, $\bar{x} = 92.2$, and $s = 10.75$, the calculated test statistic is

$$t_{\text{cal}} = \frac{92.2 - 89}{10.75/\sqrt{12}} = 1.0312.$$

Since $1.0312 < 1.795$, we cannot reject H_0 at the 5% significance level. Therefore, there is insufficient evidence to conclude that it takes longer to produce on Monday morning.

Alternatively: we find the same conclusion based on the $p-$value

$$p\text{-value} = P(T > |t_{\text{cal}}|) = P(T > 1.0312) = 0.162$$

This means that, there is a 16.2% chance of observing a sample mean of 92.2 or higher purely due to random sampling variation, assuming the null hypothesis is true ($\mu = 89$). Since p-value is larger than 0.05, then we do not reject H_0 at significance level 5%.

Problem 9.5. ***(Cardiology)*** *A recent clinical focus is on whether drugs can reduce infarct size in patients who have experienced a myocardial infarction within the last 24 hours. It is known that the average infarct size in untreated patients is 25 ($ck - g - EQ/m^2$). In contrast, among 8 patients treated with the drug, the mean infarct size is 16 with a standard deviation of 10. Is there evidence that the drug is effective in reducing infarct size?*

Solution

We set up the hypothesis test with

$$H_0 : \mu = 25 \quad \text{vs.} \quad H_1 : \mu < 25.$$

DRAFT

Here, X is normally distributed, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, with sample statistics $\bar{x} = 16$, $s = 10$, and $n = 8$. Set the significance level to $\alpha = 0.05$. From the t-distribution table, find $t_{\alpha,(n-1)} = t_{0.05,7} = 1.895$. Calculate the test statistic:

$$t_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 25}{10/\sqrt{8}} = -2.546$$

Since $|t_{\text{cal}}| = |-2.546| = 2.546$ is larger than $t_{\alpha,(n-1)} = t_{0.05,7} = 1.895$ (see Table A6 in the Appendix), we reject the null hypothesis $H_0 : \mu = 25$ at the 5% level of significance. Therefore, the evidence suggests that the mean infarct size is lower than 25.

Alternatively, the p-value is calculated as follows:

$$\begin{aligned} p\text{-value} &= P(T \geq |t_{\text{cal}}| \text{ under } H_0) \\ &= P(T \geq |-2.546|) \\ &= 0.019 \end{aligned}$$

Since p-value is less than $\alpha = 0.05$, we reject the null hypothesis $H_0 = 25$. That is, the evidence is that the mean birth-weight from this hospital is lower than the national average.

Problem 9.6. ***(Cardiovascular Disease)*** *A current area of research interest is the Family aggregation of cardiovascular risk factors in general and lipid levels in particular. Suppose the "average" cholesterol level in children is 175 mg/dL. A group of men who have died from heart disease within the past year are identified, and the cholesterol levels of their offspring are measured. Two hypotheses are considered:*

(i). The average cholesterol level of these children is 175 mg/dL.

(ii). The average cholesterol level of these children is >175 mg/dL.

Suppose the mean cholesterol level of 10 children whose fathers died from heart disease is 200 mg/dL and the sample standard deviation is 50 mg/dL. Test the hypothesis that the mean cholesterol level is higher in this group than in the general population.

Solution

To test whether the mean cholesterol level is significantly different from 175, we set up the following hypotheses:

$$H_0 : \mu = 175 \quad \text{vs.} \quad H_1 : \mu > 175.$$

Assuming that X follows a normal distribution, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, with a sample mean $\bar{x} = 200$, sample standard deviation $s = 50$, and sample size $n = 10$, we proceed with the following steps.

DRAFT

First, we set the significance level to $\alpha = 0.05$. From the t-distribution table, the critical value for $t_{9,0.95}$ is 1.833. The test statistic is calculated as follows:

$$t_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{200 - 175}{50/\sqrt{10}} = 1.58.$$

Since the calculated test statistic $|t_{\text{cal}}| = 1.58$ is less than the critical value $t_{9,0.95} = 1.833$, we do not reject the null hypothesis $H_0 : \mu = 175$ at the 5% significance level. Thus, we conclude that the mean cholesterol level of these children does not significantly differ from that of an average child.

Alternatively, we calculate the p-value as follows:

$$p\text{-value} = P(T \geq |t_{\text{cal}}| \mid H_0) = P(T \geq 1.58) = 0.074.$$

(Note: The p-value can be obtained using the function `TDIST`(1.58, 9, 1) in Excel, which yields 0.07428.)

Since the p-value is greater than $\alpha = 0.05$, we do not reject the null hypothesis $H_0 : \mu = 175$. Therefore, we conclude that the mean cholesterol level of these children does not differ significantly from that of an average child.

Problem 9.7. *Suppose the mean pulse rate in healthy adults is 72 beats per min. Research was conducted to examine the pulse rate in patients with hyperthyroidism.Twenty patients were randomly enrolled with a mean of 80 and a standard deviation of 20. Assuming that the pulse rate follows a normal distribution, is the mean pulse rate in hyperthyroidism patients different from that in healthy adults?*

Solution

To test whether the mean pulse rate differs from 72, we set up the following hypotheses:

$$H_0 : \mu = 72 \quad \text{vs.} \quad H_1 : \mu \neq 72.$$

Assuming that X follows a normal distribution, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, with a sample mean $\bar{x} = 80$, sample standard deviation $s = 20$, and sample size $n = 20$, we perform the following steps.

First, we set the significance level to $\alpha = 0.05$. For a two-tailed test with an $\alpha = 0.05$ and 19 degrees of freedom, the critical t-values can be found using a t-distribution table. The critical t-value is approximately 2.093. The test statistic is calculated as:

$$t_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{80 - 72}{20/\sqrt{20}} = 1.79.$$

Since $|t_{\text{cal}}| < 2.093$, we do not reject $H_0 : \mu = 72$. This indicates that there is not enough statistical evidence to conclude that the mean pulse rate in patients with hyperthyroidism is significantly different from that in healthy adults.

DRAFT

Alternatively, we compute the p-value for the test statistic. The p-value is given by:

$$p\text{-value} = 2P(T \geq |t_{\text{cal}}| \mid H_0) = 2P(T \geq |1.79|) = 0.0894.$$

(Note: The p-value can be obtained using the function `TDIST`(1.79, 19, 2) in Excel, which gives 0.0894.)

Since the p-value is greater than $\alpha = 0.05$, we do not reject the null hypothesis $H_0 : \mu = 72$. Thus, there is insufficient evidence to conclude that the mean pulse rate in hyperthyroidism patients differs from that in healthy adults.

□

9.5.2 Testing Equality of Two Means

Testing the equality of two means is a common procedure in hypothesis testing, used to determine if there is a significant difference between the means of two populations. This can be done using various methods depending on the nature of the samples.

9.5.3 Independent Samples T-Test

The two-sample t-test is used to compare the means of two independent groups. The hypotheses for this test are defined as follows:

- **Null Hypothesis (H_0)**: The means of the two independent samples are equal.
$$H_0 : \mu_1 = \mu_2$$
- **Alternative Hypothesis (H_1)**: The means of the two independent samples are not equal.
$$H_1 : \mu_1 \neq \mu_2$$
For a one-tailed test, the alternative hypothesis can be specified as:
  - **Right-Tailed Test:**
$$H_1 : \mu_1 > \mu_2$$
  - **Left-Tailed Test:**
$$H_1 : \mu_1 < \mu_2$$

Define the Significance Level: Choose a significance level, denoted by α. This is the probability of rejecting the null hypothesis when it is actually true. Common choices for α are 0.05, 0.01, and 0.10.

DRAFT

1. Equal Variances (Pooled T-Test)

When the variances of the two populations are assumed to be equal, the **test statistic** is calculated as follows:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Here:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means,
- s_1^2 and s_2^2 are the sample variances,
- n_1 and n_2 are the sample sizes,
- s_p^2 is the pooled variance.

For pooled T-test, the degrees of freedom is

$$\nu = n_1 + n_2 - 2.$$

Problem 9.8. ***(Hypertension)*** *A sample of eight nonpregnant, premenopausal oral contraceptive (OC) users aged 35-39 has a mean systolic blood pressure (SBP) of 132.86 mm Hg with a standard deviation of 15.34 mm Hg. In addition, a sample of 21 non-OC users from the same age group has a mean SBP of 127.44 mm Hg and a standard deviation of 18.23 mm Hg. Are the mean SBP values between these two groups equal?*

Solution

We begin by setting up the hypotheses. Let μ_1 and μ_2 represent the mean systolic blood pressures of OC users and non-OC users, respectively. The hypotheses are:

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0 \\ H_1: &\quad \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 \neq 0 \end{aligned}$$

We set the significance level α to 0.05. We consider the test statistic is

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

DRAFT

where, we assume the difference $\bar{x}_1 - \bar{x}_2$ follows a normal distribution with mean 0 and variance σ^2. The estimated variance s^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, and s_1^2 and s_2^2 are the sample variances. For our data:

$$s_p^2 = \frac{(8 - 1)(15.34)^2 + (21 - 1)(18.23)^2}{8 + 21 - 2} = \frac{8293.9}{27} = 307.18$$

Based on the sample information, the realized value of T is

$$t_{\text{cal}} = \frac{132.86 - 127.44}{\sqrt{307.18\left(\frac{1}{8} + \frac{1}{21}\right)}} = 0.74$$

For a two-tailed test at the 0.05 significance level with degrees of freedom 27, the critical t-value of the right-side is approximately 2.052.

Since $|t_{\text{cal}}| = 0.74 < 2.052$, the we do not reject H_0. This indicates that there is not enough statistical evidence to conclude that the mean systolic blood pressure between oral contraceptive users and non-users is significantly different.

Alternatively, the p-value is calculated as:

$$p\text{-value} = 2P(T \geq |t_{\text{cal}}| \mid H_0) = 2P(T \geq |0.74|) = 0.46$$

Since the p-value of 0.46 is greater than the significance level $\alpha = 0.05$, we do not reject the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$. Therefore, we conclude that there is no significant difference in mean systolic blood pressure between OC users and non-OC users in the specified age group.

□

2. Unequal Variances (Welch's T-Test)

When the variances of the two populations are not assumed to be equal, use Welch's T-test. The **test statistic** T is as follows:

$$T = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For the Welch's T-test, the degrees of freedom is

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Decision Rule: Compare the calculated t-statistic to the critical value:

- **Reject** H_0 if $|t_{\text{cal}}| \geq$ critical value or if the p-value $\leq \alpha$.
- **Fail to Reject** H_0 if $|t_{\text{cal}}| >$ critical value or if the p-value $> \alpha$.

Problem 9.9. *A researcher is investigating whether two different training programs lead to different levels of improvement in employee productivity. A random sample of 15 employees who completed Training Program A showed a mean improvement score of 78 points with a standard deviation of 12 points. A random sample of 20 employees who completed Training Program B showed a mean improvement score of 85 points with a standard deviation of 22 points. Assuming the population variances are not equal, test whether there is a significant difference in mean improvement scores between the two programs at the 5% significance level.*

Solution:

To test whether there is a significant difference in mean improvement scores between the two training programs when the population variances are not equal, we use Welch's T-test. We have data from Training Program A with $n_1 = 15$, $\bar{x}_1 = 78$, and $s_1 = 12$, and from Training Program B with $n_2 = 20$, $\bar{x}_2 = 85$, and $s_2 = 22$.

For testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at significance level $\alpha = 0.05$, Welch's T-statistic is

$$t_{\text{cal}} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{85 - 78}{\sqrt{\frac{144}{15} + \frac{484}{20}}} = \frac{7}{\sqrt{9.6 + 24.2}} = \frac{7}{\sqrt{33.8}} = \frac{7}{5.814} = 1.204$$

The degrees of freedom for Welch's t-test are approximated using the Welch-Satterthwaite equation:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{(33.8)^2}{\frac{(9.6)^2}{14} + \frac{(24.2)^2}{19}}$$
$$= \frac{1142.44}{6.583 + 30.823} = 30.54$$

We use $\nu \approx 30$ for practical purposes. For a two-tailed test with $\alpha = 0.05$ and $\nu = 30$, the critical value is $t_{0.025,30} = 2.042$. Since $|t_{\text{cal}}| = 1.204 < 2.042$, we fail to reject the null hypothesis at the 5% significance level. □

9.5.4 Paired T-Test

The paired t-test is used to determine whether there is a significant difference between the means of two related groups. This test is commonly used in experiments where measurements are taken on the same subjects under two different conditions or at two different times. The hypotheses for the paired T-test are defined as follows:

DRAFT

- **Null Hypothesis (H_0)**: The mean difference between paired observations is zero.
$$H_0 : \mu_d = 0$$

- **Alternative Hypothesis (H_1)**: The mean difference between paired observations is not zero.
$$H_1 : \mu_d \neq 0$$
For a one-tailed test, the alternative hypothesis can be specified as:
  - **Right-Tailed Test:**
$$H_1 : \mu_d > 0$$
  - **Left-Tailed Test:**
$$H_1 : \mu_d < 0$$

Test Statistic

Calculate the test statistic t using the formula:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean difference, s_d is the standard deviation of the differences, and n is the number of pairs.

Critical Value and Decision Rule:

- Determine the critical value from the t-distribution table based on α and degrees of freedom ($df = n - 1$).
- For a two-tailed test, reject H_0 if $|t|$ is greater than the critical value. For a one-tailed test, reject H_0 if t falls in the direction specified by H_a.

Problem 9.10. *A nutritionist wants to evaluate the effectiveness of a new dietary supplement in reducing cholesterol levels. A sample of 10 participants had their cholesterol levels measured before and after taking the supplement for a month. The cholesterol levels (in mg/dL) before and after the treatment are provided as follows:*

DRAFT

Before	***After***
220	*210*
240	*230*
215	*205*
230	*215*
250	*240*
210	*200*
225	*210*
240	*220*
235	*225*
245	*230*

Test if the dietary supplement had a significant effect on cholesterol levels at the 5% significance level.

Solution

To determine whether the dietary supplement significantly affects cholesterol levels, we use a paired t-test.

Hypotheses:

- **Null Hypothesis (H_0)**: There is no difference in cholesterol levels before and after the treatment, i.e., $\mu_d = 0$.
- **Alternative Hypothesis (H_1)**: There is a significant difference in cholesterol levels, i.e., $\mu_d \neq 0$.

where μ_d is the mean difference in cholesterol levels before and after the treatment.

First, we calculate the differences between the before and after measurements for each participant:

$$\text{Differences} = [10, 10, 10, 15, 10, 10, 15, 20, 10, 15]$$

DRAFT

i	**Before**	**After**	**Difference** (d_i)
1	220	210	10
2	240	230	10
3	215	205	10
4	230	215	15
5	250	240	10
6	210	200	10
7	225	210	15
8	240	220	20
9	235	225	10
10	245	230	15

Next, we compute the mean difference ($\bar{d}$) and the standard deviation of the differences (s_d):

$$\bar{d} = \frac{10 + 10 + 10 + 15 + 10 + 10 + 15 + 20 + 10 + 15}{10} = 12.5$$

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{(10 - 12.5)^2 + (10 - 12.5)^2 + \cdots + (15 - 12.5)^2}{9}} = 3.5355$$

The test statistic is calculated using:

$$t_{\text{cal}} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{12.5}{3.5355/\sqrt{10}} = \frac{12.5}{0.91} \approx 11.1803$$

With $n - 1 = 9$ degrees of freedom, we compare the calculated t-value to the critical value from the t-distribution table at the 5% significance level. For a two-tailed test with 9 degrees of freedom, the critical value is approximately 2.262.

Since 11.1803 exceeds 2.262, we reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the dietary supplement has a significant effect on reducing cholesterol levels. □

9.6 Testing Equality of Several Means

9.6.1 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to compare the means of multiple groups and assess whether there are statistically significant

DRAFT

differences between them. It is particularly useful when comparing **more than two independent groups** or treatments.

For instance, to evaluate the effectiveness of different diabetes medications, researchers design an experiment to explore the relationship between the type of medication and the resulting blood sugar levels. The study involves a sample population, which is divided into several groups, each receiving a specific medication over a trial period. After the trial, blood sugar levels are measured for each participant. The mean blood sugar level is then calculated for each group. ANOVA is used to compare these group means to determine if there are significant differences, or if the means are statistically similar.

This method is also known as Fisher's analysis of variance, emphasizing its capacity to examine how a categorical variable with multiple levels affects a continuous variable. The application of ANOVA is determined by the research design. Typically, ANOVAs are employed in three main forms: one-way ANOVA, two-way ANOVA, and N-way ANOVA.

Data Layout for One-Way ANOVA

Suppose we have k groups, each corresponding to a different treatment or category. Let,

- x_{ij} denote the j-th observation in the i-th group, where:
  - $i = 1, 2, \ldots, k$, indicating the group number
  - $j = 1, 2, \ldots, n_i$, indicating the observation number within group i
- n_i denote the sample size of group i
- $\bar{x}_i$ denote the mean of the i-th group
- s_i^2 denote the variance of the i-th group

The data are organized as follows:

Group 1	**Group 2**	$\cdots$	**Group** k
x_{11}	x_{21}	$\cdots$	x_{k1}
x_{12}	x_{22}	$\cdots$	x_{k2}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_{1n_1}	x_{2n_2}	$\cdots$	x_{kn_k}
$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$
s_1^2	s_2^2	$\cdots$	s_k^2

DRAFT

State the Hypotheses

- **Null Hypothesis (H_0)**: All group means are equal.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- **Alternative Hypothesis (H_1)**: At least one group mean is different.

$$H_1 : \text{At least one } \mu_i \text{ differs from the others}$$

Collect and Summarize the Data

Organize the data into groups and compute the following for each group:

- $\bar{x}_i$: Mean of the i-th group
- s_i^2: Sample variance of the i-th group
- n_i: Number of observations in the i-th group

Calculate the overall (grand) mean:

$$\bar{x}_{\text{overall}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

Compute the ANOVA Table

To conduct a one-way ANOVA, compute the following components and summarize the results in the ANOVA table (see Table 9.2).

- **Total number of observations:** $n = \sum_{i=1}^{k} n_i$
- **Sum of Squares Between Groups (SSB):**

$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x}_{\text{overall}})^2$$

- **Sum of Squares Within Groups (SSW):**

$$\text{SSW} = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

- **Total Sum of Squares (SST):**

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\text{overall}})^2 = \text{SSB} + \text{SSW}$$

DRAFT

Table 9.2: ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Statistic
Between Groups (SSB)	SSB	$k-1$	$\text{MSB} = \frac{\text{SSB}}{k-1}$	$F = \frac{\text{MSB}}{\text{MSW}}$
Within Groups (SSW)	SSW	$n-k$	$\text{MSW} = \frac{\text{SSW}}{n-k}$	
Total (SST)	SST	$n-1$		

Determine the Critical Value or p-Value

Find the critical value from the F-distribution table using degrees of freedom $\text{df}_{\text{between}} = k-1$ and $\text{df}_{\text{within}} = n-k$ at the desired significance level α. The F-distribution table is presented in Table A8 of the Appendix. Alternatively, calculate the p-value associated with the observed F-statistic.

Make a Decision

- **If using the critical value:**
  - If $F > F_{\text{critical}}$, reject the null hypothesis H_0.
  - If $F \leq F_{\text{critical}}$, do not reject H_0.
- **If using the p-value:**
  - If $p \leq \alpha$, reject the null hypothesis H_0.
  - If $p > \alpha$, do not reject the null hypothesis H_0.

Interpret the Results

- If the null hypothesis is rejected, we conclude that there is a statistically significant difference in means among the groups.
- If the null hypothesis is not rejected, we conclude that the group means are statistically similar.

Problem 9.11. *Suppose we want to compare the effectiveness of three different blood pressure medications: Medication A, Medication B, and Medication C. We have following data on the systolic blood pressure of patients who received each of the three medications.*

- ***Medication A**: 140, 145, 148, 142, 144*
- ***Medication B**: 130, 133, 135, 138, 132*

DRAFT

- ***Medication C**: 125, 128, 124, 127, 126*

Test if there are significant differences in the mean blood pressure levels among these groups at the 5% significance level.

State the Hypotheses

- **Null Hypothesis (H_0)**: The mean systolic blood pressure is the same across all three groups (Medication A, Medication B, Medication C).

$$H_0 : \mu_A = \mu_B = \mu_C$$

- **Alternative Hypothesis (H_1)**: At least one of the group means is different.

$$H_1 : \text{At least one } \mu_i \text{ differs from the others.}$$

Collect and Summarize the Data

The data collected for the three groups are as follows:

	Medication A	Medication B	Medication C
	140	130	125
	145	133	128
	148	135	124
	142	138	127
	144	132	126
Mean ($\bar{x}_i$)	143.8	133.6	126
Variance (s_i^2)	9.2	9.3	2.5

We now calculate the overall mean ($\bar{x}_{\text{overall}}$):

$$\bar{x}_{\text{overall}} = \frac{5 \times 143.8 + 5 \times 133.6 + 5 \times 126}{5 + 5 + 5} = 134.47$$

Compute the ANOVA Table

Sum of Squares Between Groups (SSB):

$$\begin{aligned}\text{SSB} &= n_1(\bar{x}_1 - \bar{x}_{\text{overall}})^2 + n_2(\bar{x}_2 - \bar{x}_{\text{overall}})^2 + n_3(\bar{x}_3 - \bar{x}_{\text{overall}})^2 \\ &= 5 \times (143.8 - 134.47)^2 + 5 \times (133.6 - 134.47)^2 + 5 \times (126 - 134.47)^2 \\ &= 797.73\end{aligned}$$

DRAFT

Sum of Squares Within Groups (SSW):

$$\begin{aligned} \text{SSW} &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 \\ &= 4 \times 9.2 + 4 \times 9.3 + 4 \times 2.5 \\ &= 84 \end{aligned}$$

Degrees of Freedom:

- Between Groups: $\text{df}_{\text{between}} = k - 1 = 3 - 1 = 2$
- Within Groups: $\text{df}_{\text{within}} = n - k = 15 - 3 = 12$

Mean Squares:

- Mean Square Between Groups (MSB):

$$\text{MSB} = \frac{\text{SSB}}{\text{df}_{\text{between}}} = \frac{797.73}{2} = 398.87$$

- Mean Square Within Groups (MSW):

$$\text{MSW} = \frac{\text{SSW}}{\text{df}_{\text{within}}} = \frac{84}{12} = 7$$

***F*-Statistic**:

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{398.87}{7} = 56.98$$

Table 9.3: ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Statistic
Between Groups	797.73	2	398.87	56.98
Within Groups	84	12	7	-
Total	881.7333	14	-	-

Determine the Critical Value or p-Value

Using an F-distribution table, find the critical value for $\text{df}_{\text{between}} = 2$ and $\text{df}_{\text{within}} = 12$ at a significance level $\alpha = 0.05$. The critical value for $F(2, 12)$ at $\alpha = 0.05$ is approximately 3.89.

Alternatively, using statistical software, we can calculate the p-value associated with the F-statistic of 56.98, which is very small (much less than 0.05).

DRAFT

Make a Decision

- **Using the critical value**: Since $F = 56.98$ is greater than the critical value of 3.89, we reject the null hypothesis.
- **Using the p-value**: Since the p-value is less than 0.05, we also reject the null hypothesis.

Problem 9.12. *A company wants to determine if there are significant differences in the average productivity of employees working under three different types of work environments. The productivity scores (in units produced per hour) are collected from three groups of employees in different environments: Environment A, Environment B, and Environment C. The data collected are:*

Environment A	***Environment B***	***Environment C***
30	*45*	*55*
32	*50*	*53*
29	*47*	*52*
31	*46*	*58*
33	*48*	*60*

Test if there are significant differences in the mean productivity across the three environments at the 5% significance level.

Solution

Calculate Means and Overall Mean

Table 9.4: Means of Productivity Scores

Environment	**Mean**
Environment A	$\bar{x}_A = \frac{30+32+29+31+33}{5} = 31.00$
Environment B	$\bar{x}_B = \frac{45+50+47+46+48}{5} = 47.20$
Environment C	$\bar{x}_C = \frac{55+53+52+58+60}{5} = 55.60$
Overall Mean	$\bar{x}_{\text{overall}} = \frac{31+47.20+55.60}{3} = 44.6$

DRAFT

	Environment A	Environment B	Environment C
	30	45	55
	32	50	53
	29	47	52
	31	46	58
	33	48	60
Mean ($\bar{x}_i$)	31	47.2	55.6
Variance (s_i^2)	2.5	3.7	11.3

Sum of Squares Between Groups (SSB)

$$\begin{aligned}\text{SSB} &= n\left[(\bar{x}_A - \bar{x}_{\text{overall}})^2 + (\bar{x}_B - \bar{x}_{\text{overall}})^2 + (\bar{x}_C - \bar{x}_{\text{overall}})^2\right] \\ &= 5\left[(31 - 44.6)^2 + (47.20 - 44.6)^2 + (55.60 - 44.6)^2\right] \\ &= 1563.6\end{aligned}$$

Sum of Squares Within Groups (SSW)

Calculate the sum of squares within each group. For simplicity, assume the following values for demonstration:

$$\text{SSW} = (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2 = 70$$

Mean Squares

$$\text{MSB} = \frac{\text{SSB}}{k - 1} = \frac{1563.6}{3 - 1} = 781.8$$

$$\text{MSW} = \frac{\text{SSW}}{n - k} = \frac{70}{15 - 3} = 5.83$$

***F*-Statistic**

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{781.8}{5.83} \approx 134.02$$

Conclusion

Compare the calculated F-value to the critical value from the F-distribution table with $k - 1$ and $n - k$ degrees of freedom at $\alpha = 0.05$.

Assuming the critical value from the F-table is approximately 3.89:

- Since the calculated F-value (134.02) is greater than the critical value (3.89), we reject the null hypothesis.
- There is a significant difference in the mean productivity among the three environments.

DRAFT

Table 9.5: ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Statistic
Between Groups	1563.6	2	781.8	134.02
Within Groups	70	12	5.83	-
Total	1633.6	14	-	-

Python Code

To solve Problem 9.12, the following Python code performs a one-way ANOVA test:

```
import numpy as np
import pandas as pd
from scipy import stats

# Step 1: Define the data for each environment group
env_a = np.array([30, 32, 29, 31, 33])
env_b = np.array([45, 50, 47, 46, 48])
env_c = np.array([55, 53, 52, 58, 60])

# Step 2: (Optional) Create a DataFrame for inspection
df = pd.DataFrame({
  'Environment A': env_a,
  'Environment B': env_b,
  'Environment C': env_c
})
print("Data:\n", df)

# Step 3: Calculate group means and variances
means = [np.mean(env_a), np.mean(env_b), np.mean(env_c)]
variances = [np.var(env_a, ddof=1), np.var(env_b, ddof=1),
   np.var(env_c, ddof=1)]
print("\nGroup Means:", means)
print("Group Variances:", variances)

# Step 4: Perform one-way ANOVA using scipy.stats.f_oneway
f_stat, p_value = stats.f_oneway(env_a, env_b, env_c)

print("\nANOVA Result:")
print(f"F-statistic = {f_stat:.4f}")
print(f"P-value = {p_value:.4f}")

# Step 5: Interpretation at alpha = 0.05 significance
  level
```

```
alpha = 0.05
if p_value < alpha:
print("\nConclusion: Reject the null hypothesis.")
print("There is a significant difference among the group
  means.")
else:
print("\nConclusion: Do not reject the null hypothesis.")
print("No significant difference among the group means.")
```

9.7 Tests for Variance

After examining hypothesis tests for population means, we now turn our attention to testing claims about population variances. Variance tests are crucial in many applications where the variability or consistency of a process is of primary interest. For instance, quality control engineers might need to verify that manufacturing processes maintain consistent precision, or financial analysts might assess whether investment portfolios exhibit acceptable levels of risk.

9.7.1 One-Sample Variance Test

The one-sample variance test (also called the chi-square test for variance) is used to test hypotheses about a single population variance σ^2 when the population is normally distributed.

Assumptions

For the one-sample variance test to be valid, the following assumptions must be satisfied:

(i). The sample is randomly selected from the population.

(ii). The population from which the sample is drawn follows a normal distribution.

(iii). The observations are independent.

Note that the variance test is particularly sensitive to violations of the normality assumption, much more so than the mean test. When the population distribution is significantly non-normal, alternative methods such as bootstrap procedures should be considered.

Test Statistic

Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution with unknown variance σ^2. The sample variance is defined as:

DRAFT

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{9.1}$$

To test hypotheses about σ^2, we use the chi-square test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{9.2}$$

where σ_0^2 is the hypothesized value of the population variance, and n is the sample size.

Under the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$, this test statistic follows a chi-square distribution with $\nu = n - 1$ degrees of freedom, denoted as χ^2_{n-1}.

Hypothesis Testing Framework

The general framework for testing variance can take one of three forms:

1. **Two-tailed test:**

$$H_0 : \sigma^2 = \sigma_0^2 \tag{9.3}$$
$$H_1 : \sigma^2 \neq \sigma_0^2 \tag{9.4}$$

   Rejection region: $\chi^2 < \chi^2_{\alpha/2,n-1}$ or $\chi^2 > \chi^2_{1-\alpha/2,n-1}$

2. **Right-tailed test:**

$$H_0 : \sigma^2 \leq \sigma_0^2 \tag{9.5}$$
$$H_1 : \sigma^2 > \sigma_0^2 \tag{9.6}$$

   Rejection region: $\chi^2 > \chi^2_{1-\alpha,n-1}$

3. **Left-tailed test:**

$$H_0 : \sigma^2 \geq \sigma_0^2 \tag{9.7}$$
$$H_1 : \sigma^2 < \sigma_0^2 \tag{9.8}$$

   Rejection region: $\chi^2 < \chi^2_{\alpha,n-1}$

where α is the significance level and $\chi^2_{\alpha,n-1}$ represents the critical value from the chi-square distribution with $n - 1$ degrees of freedom.

Problem 9.13. *A pharmaceutical company manufactures tablets that should have a weight variance not exceeding 0.04 grams^2. A quality control inspector randomly selects 25 tablets and finds a sample variance of $s^2 = 0.053$ grams^2. At the $\alpha = 0.05$ significance level, test whether the variance exceeds the specification.*

DRAFT

Solution

Step 1: State the hypotheses

$$H_0 : \sigma^2 \leq 0.04$$
$$H_1 : \sigma^2 > 0.04$$

Step 2: Determine the significance level and critical value

With $\alpha = 0.05$ and $n - 1 = 24$ degrees of freedom, the critical value is $\chi^2_{0.95,24} = 36.415$.

Step 3: Calculate the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 \times 0.053}{0.04} = \frac{1.272}{0.04} = 31.8$$

Step 4: Make a decision

Since $\chi^2 = 31.8 < 36.415$, we fail to reject H_0 at the 0.05 significance level.

Step 5: Conclusion

There is insufficient evidence to conclude that the variance of tablet weights exceeds 0.04 grams^2 at the 5% significance level. The manufacturing process appears to meet the variance specification. □

9.7.2 Two-Sample Variance Test (F-Test)

The F-test for equality of variances is used to compare the variances of two independent populations. This test is commonly performed before conducting a two-sample t-test to determine whether to use the pooled or unpooled variance version.

Assumptions

(i). Both samples are randomly and independently selected from their respective populations.

(ii). Both populations follow normal distributions.

(iii). The observations within each sample are independent.

Like the one-sample variance test, the F-test is sensitive to departures from normality. Levene's test or the Brown-Forsythe test are robust alternatives when normality cannot be assumed.

Test Statistic

Let s_1^2 and s_2^2 be the sample variances from two independent samples of sizes n_1 and n_2, respectively. The F-statistic is defined as:

$$F = \frac{s_1^2}{s_2^2} \tag{9.9}$$

DRAFT

By convention, we typically place the larger sample variance in the numerator, so that $F \geq 1$.

Under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, this statistic follows an F-distribution with $\nu_1 = n_1 - 1$ degrees of freedom in the numerator and $\nu_2 = n_2 - 1$ degrees of freedom in the denominator, denoted as F_{n_1-1,n_2-1}.

Hypothesis Testing Framework

1. **Two-tailed test:**

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{9.10}$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{9.11}$$

   Rejection region: $F > F_{\alpha/2,n_1-1,n_2-1}$ (when $s_1^2 > s_2^2$)

2. **Right-tailed test:**

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \tag{9.12}$$
$$H_1 : \sigma_1^2 > \sigma_2^2 \tag{9.13}$$

   Rejection region: $F > F_{\alpha,n_1-1,n_2-1}$

3. **Left-tailed test:**

$$H_0 : \sigma_1^2 \geq \sigma_2^2 \tag{9.14}$$
$$H_1 : \sigma_1^2 < \sigma_2^2 \tag{9.15}$$

   Rejection region: $F < F_{1-\alpha,n_1-1,n_2-1}$ or equivalently $F > F_{\alpha,n_2-1,n_1-1}$ with swapped variances

Problem 9.14. *A manufacturing company wants to compare the consistency of two production lines. From Line A, a random sample of $n_1 = 16$ items yields a variance of $s_1^2 = 12.5$. From Line B, a random sample of $n_2 = 21$ items yields a variance of $s_2^2 = 6.8$. Test at the $\alpha = 0.10$ significance level whether the variances differ significantly.*

Solution

To compare the consistency of the two production lines, we test the hypotheses $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$ using the F-test at significance level $\alpha = 0.10$. Placing the larger variance in the numerator, the test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{12.5}{6.8} = 1.838$$

with degrees of freedom $\nu_1 = 15$ and $\nu_2 = 20$. For a two-tailed test with $\alpha = 0.10$, we use $\alpha/2 = 0.05$ in each tail, giving a critical value of $F_{0.05,15,20} =$

DRAFT

2.20. Since $F = 1.838 < 2.20$, we fail to reject the null hypothesis at the 10% significance level. There is insufficient evidence to conclude that the variances of the two production lines differ significantly. The production lines appear to have similar variability.

□

Problem 9.15. *Refer to the Example 9.8. A sample of eight nonpregnant, premenopausal oral contraceptive (OC) users aged 35-39 has a mean systolic blood pressure (SBP) of 132.86 mm Hg with a standard deviation of 15.34 mm Hg. In addition, a sample of 21 non-OC users from the same age group has a mean SBP of 127.44 mm Hg and a standard deviation of 18.23 mm Hg. Are the variance SBP values between these two groups equal?*

Solution:

To determine whether the variance of systolic blood pressure differs between oral contraceptive users and non-users, we conduct a two-sample F-test at the $\alpha = 0.05$ significance level. The hypotheses are $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$, where σ_1^2 and σ_2^2 represent the population variances for OC users and non-OC users, respectively.

The sample variances are computed as $s_1^2 = (15.34)^2 = 235.316$ mm Hg^2 and $s_2^2 = (18.23)^2 = 332.333$ mm Hg^2. Placing the larger variance in the numerator, the F-statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{332.333}{235.316} = 1.412$$

with degrees of freedom $\nu_1 = 20$ and $\nu_2 = 7$. For a two-tailed test at $\alpha = 0.05$, the critical value is $F_{0.025,20,7} = 4.47$. Since $F = 1.412 < 4.47$, we fail to reject the null hypothesis.

We can also construct a 95% confidence interval for the variance ratio σ_2^2/σ_1^2:

$$\left(\frac{1.412}{4.47}, 1.412 \times 3.44\right) = (0.316, 4.857)$$

Since this interval contains 1, it confirms that there is insufficient evidence to conclude that the variances differ between the two groups. The finding that the variances are not significantly different has important implications for further analysis. When comparing mean SBP between the two groups, we can use the pooled two-sample t-test, which assumes equal variances and is generally more powerful than the unpooled version when this assumption is reasonable. □

Problem 9.16. *Refer to the example* **??**. *A data scientist wants to determine if there is a significant difference in the mean sales figures before and after implementing a new marketing strategy. The sales data before the strategy (sample size = 25) has a mean of $5000 with a standard deviation of $800, and the sales data after the strategy (sample size = 30) has a mean of $5300 with a standard deviation of $850. Test whether the mean sales of both groups are equal at a 10% significance level.*

DRAFT

Solution: To test whether there is a significant difference in mean sales before and after implementing the new marketing strategy, we conduct a two-sample t-test. We have sales data before the strategy with $n_1 = 25$, $\bar{x}_1 = \$5000$, and $s_1 = \$800$, and sales data after the strategy with $n_2 = 30$, $\bar{x}_2 = \$5300$, and $s_2 = \$850$. The hypotheses are $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at significance level $\alpha = 0.10$.

Before conducting the test, we first check whether the population variances can be assumed equal using an F-test. The sample variances are $s_1^2 = (800)^2 = 640{,}000$ and $s_2^2 = (850)^2 = 722{,}500$. The F-statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{722{,}500}{640{,}000} = 1.129$$

with degrees of freedom $\nu_1 = 29$ and $\nu_2 = 24$. At $\alpha = 0.10$, the critical value is $F_{0.05,29,24} = 2.01$. Since $F = 1.129 < 2.01$, we fail to reject the hypothesis of equal variances. Therefore, we proceed with the pooled two-sample t-test.

The pooled variance is calculated as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{24(640{,}000) + 29(722{,}500)}{53} = 685330.2$$

The pooled standard deviation is $s_p = \sqrt{685{,}330.2} = 827.85$. The test statistic is

$$t_{\text{cal}} = \frac{\bar{x}_2 - \bar{x}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{5300 - 5000}{827.85\sqrt{\frac{1}{25} + \frac{1}{30}}} = \frac{300}{827.85\sqrt{0.0733}} = \frac{300}{224.21} = 1.338$$

For a two-tailed test with $\alpha = 0.10$ and degrees of freedom $df = n_1 + n_2 - 2 = 53$, the critical value is approximately $t_{0.05,53} = 1.674$. Since $|t_{\text{cal}}| = 1.338 < 1.674$, we fail to reject the null hypothesis at the 10% significance level. □

Problem 9.17. *A researcher is investigating whether two different training programs lead to different levels of improvement in employee productivity. A random sample of 15 employees who completed Training Program A showed a mean improvement score of 78 points with a standard deviation of 12 points. A random sample of 20 employees who completed Training Program B showed a mean improvement score of 85 points with a standard deviation of 22 points. Test whether there is a significant difference in mean improvement scores between the two programs at the 5% significance level.*

Solution: To test whether there is a significant difference in mean improvement scores between the two training programs, we first examine whether the population variances can be assumed equal. We have data from Training Program A with $n_1 = 15$, $\bar{x}_1 = 78$, and $s_1 = 12$, and from Training Program B with $n_2 = 20$, $\bar{x}_2 = 85$, and $s_2 = 22$.

We begin by testing the equality of variances using an F-test. The sample variances are $s_1^2 = (12)^2 = 144$ and $s_2^2 = (22)^2 = 484$. The F-statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{484}{144} = 3.361$$

DRAFT

with degrees of freedom $\nu_1 = 19$ and $\nu_2 = 14$. For a two-tailed test at $\alpha = 0.05$, the critical value is $F_{0.025,19,14} = 3.03$. Since $F = 3.361 > 3.03$, we reject the hypothesis of equal variances. Therefore, we must use Welch's t-test (the unpooled two-sample t-test), which does not assume equal variances.

For testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ at significance level $\alpha = 0.05$, Welch's t-statistic is

$$t_{\text{cal}} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{85 - 78}{\sqrt{\frac{144}{15} + \frac{484}{20}}} = \frac{7}{\sqrt{9.6 + 24.2}} = \frac{7}{\sqrt{33.8}} = \frac{7}{5.814} = 1.204$$

The degrees of freedom for Welch's t-test are approximated using the Welch-Satterthwaite equation:

$$\begin{aligned} df &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{(33.8)^2}{\frac{(9.6)^2}{14} + \frac{(24.2)^2}{19}} \\ &= \frac{1142.44}{\frac{92.16}{14} + \frac{585.64}{19}} = \frac{1142.44}{6.583 + 30.823} \\ &= \frac{1142.44}{37.406} = 30.54 \end{aligned}$$

We use $df \approx 30$ for practical purposes. For a two-tailed test with $\alpha = 0.05$ and $df = 30$, the critical value is $t_{0.025,30} = 2.042$. Since $|t_{\text{cal}}| = 1.204 < 2.042$, we fail to reject the null hypothesis at the 5% significance level. While Training Program B shows a higher average improvement of 7 points, this difference is not statistically significant given the substantial variability in the data, particularly for Program B. The researcher should note that the higher variability in Program B's results ($s_2 = 22$ versus $s_1 = 12$) suggests that this program may produce more variable outcomes across different employees, which is an important practical consideration beyond the mean comparison. □

9.7.3 Confidence Intervals for Variance

In addition to hypothesis testing, we can construct confidence intervals for population variances.

Confidence Interval for a Single Variance

A $(1 - \alpha) \times 100\%$ confidence interval for σ^2 is given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \right) \tag{9.16}$$

The corresponding confidence interval for the standard deviation σ is:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}} \right) \tag{9.17}$$

DRAFT

Confidence Interval for the Ratio of Two Variances

A $(1-\alpha)\times 100\%$ confidence interval for the ratio σ_1^2/σ_2^2 is given by:

$$\left(\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2,n_2-1,n_1-1}\right) \tag{9.18}$$

If this interval contains 1, we fail to reject the null hypothesis that the two variances are equal.

9.7.4 Tests for Homogeneity of Variance Across Multiple Groups

When comparing variances across more than two populations, conducting multiple pairwise F-tests is inappropriate due to the inflation of Type I error rates. Instead, we use tests specifically designed for multiple group comparisons. The two most common tests are Bartlett's test and Levene's test.

Bartlett's Test

Bartlett's test is used to test the null hypothesis that all k populations have equal variances. It is the extension of the F-test to multiple groups.

Assumptions:

(i). All samples are randomly and independently selected.

(ii). All populations follow normal distributions.

(iii). Observations within each sample are independent.

Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \tag{9.19}$$

$$H_1 : \text{At least one variance differs from the others} \tag{9.20}$$

Test Statistic:

Let s_i^2 be the sample variance of the i-th group with sample size n_i, for $i = 1, 2, \ldots, k$. The pooled variance is:

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{\sum_{i=1}^{k}(n_i-1)} \tag{9.21}$$

Bartlett's test statistic is:

$$\chi_B^2 = \frac{1}{C}\left[N-k)\ln(s_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(s_i^2)\right] \tag{9.22}$$

where $N = \sum_{i=1}^{k} n_i$ is the total sample size, and the correction factor C is:

DRAFT

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N-k} \right] \tag{9.23}$$

Under H_0, the test statistic χ^2_B approximately follows a chi-square distribution with $k-1$ degrees of freedom. We reject H_0 if $\chi^2_B > \chi^2_{\alpha,k-1}$.
Important Note: Bartlett's test is highly sensitive to departures from normality. Even slight deviations from normality can lead to unreliable results. When normality is questionable, Levene's test is preferred.

Levene's Test

Levene's test is a robust alternative to Bartlett's test that is less sensitive to departures from normality. It tests the equality of variances by comparing the variability of observations from their group means.

Assumptions:

(i). All samples are randomly and independently selected.

(ii). Observations within each sample are independent.

(iii). (No strict normality assumption required)

Procedure:
For each observation x_{ij} (the j-th observation in group i), compute the absolute deviation from the group mean (or median):

$$z_{ij} = |x_{ij} - \bar{x}_i| \tag{9.24}$$

or, for the more robust version (Brown-Forsythe test):

$$z_{ij} = |x_{ij} - \tilde{x}_i| \tag{9.25}$$

where $\bar{x}_i$ is the mean of group i and $\tilde{x}_i$ is the median of group i.
The test then performs a one-way ANOVA on these z_{ij} values. The test statistic is:

$$F_L = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} n_i (\bar{z}_i - \bar{z})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2} \tag{9.26}$$

where:

- $\bar{z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}$ is the mean of the z values in group i
- $\bar{z} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij}$ is the overall mean of all z values

Under H_0, F_L follows an F-distribution with $(k-1)$ and $(N-k)$ degrees of freedom. We reject H_0 if $F_L > F_{\alpha,k-1,N-k}$.

DRAFT

Comparison of Bartlett's and Levene's Tests

Table 9.6: Comparison of Bartlett's and Levene's tests for homogeneity of variance

Criterion	**Bartlett's Test**	**Levene's Test**
Normality assumption	Requires strict normality	Robust to non-normality
Sensitivity	Highly sensitive to outliers and non-normality	Less sensitive to outliers
Power (normal data)	More powerful when data are normal	Slightly less powerful
Recommended use	Only when normality is verified	General use, especially with non-normal or unknown distributions

Problem 9.18. *An educational researcher wants to compare the consistency of student performance across four different teaching methods. Random samples of students from each method yield the following test score data:*

Method	**Sample Size**	**Sample Variance**	**Sample Mean**
Method 1	$n_1 = 15$	$s_1^2 = 64$	$\bar{X}_1 = 78$
Method 2	$n_2 = 18$	$s_2^2 = 81$	$\bar{X}_2 = 82$
Method 3	$n_3 = 12$	$s_3^2 = 49$	$\bar{X}_3 = 75$
Method 4	$n_4 = 20$	$s_4^2 = 121$	$\bar{X}_4 = 80$

Test at $\alpha = 0.05$ whether the variances are equal using Bartlett's test.

Solution

Step 1: State the hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$
$$H_1 : \text{At least one variance differs}$$

Step 2: Calculate the pooled variance
Total degrees of freedom: $N - k = (15 + 18 + 12 + 20) - 4 = 61$

DRAFT

$$
\begin{aligned}
s_p^2 &= \frac{(15-1)(64)+(18-1)(81)+(12-1)(49)+(20-1)(121)}{61} \\
&= \frac{14(64)+17(81)+11(49)+19(121)}{61} \\
&= \frac{896+1377+539+2299}{61} \\
&= \frac{5111}{61} = 83.787
\end{aligned}
$$

Step 3: Calculate the correction factor

$$
\begin{aligned}
C &= 1+\frac{1}{3(4-1)}\left[\frac{1}{14}+\frac{1}{17}+\frac{1}{11}+\frac{1}{19}-\frac{1}{61}\right] \\
&= 1+\frac{1}{9}\left[0.0714+0.0588+0.0909+0.0526-0.0164\right] \\
&= 1+\frac{1}{9}(0.2573) \\
&= 1.0286
\end{aligned}
$$

Step 4: Calculate Bartlett's test statistic

$$
\begin{aligned}
\chi_B^2 &= \frac{1}{1.0286}\left[61\ln(83.787)-14\ln(64)-17\ln(81)-11\ln(49)-19\ln(121)\right] \\
&= \frac{1}{1.0286}\left[61(4.428)-14(4.159)-17(4.394)-11(3.892)-19(4.796)\right] \\
&= \frac{1}{1.0286}\left[270.108-58.226-74.698-42.812-91.124\right] \\
&= \frac{3.248}{1.0286} = 3.158
\end{aligned}
$$

Step 5: Determine the critical value and make a decision

With $k-1=3$ degrees of freedom and $\alpha=0.05$, the critical value is $\chi^2_{0.05,3}=7.815$.

Since $\chi_B^2 = 3.158 < 7.815$, we fail to reject H_0.

Step 6: Conclusion

At the 5% significance level, there is insufficient evidence to conclude that the variances of test scores differ across the four teaching methods. The assumption of homogeneity of variance is reasonable for these data. □

Post-Hoc Analysis

If the null hypothesis is rejected in either Bartlett's or Levene's test, indicating that at least one variance differs, post-hoc pairwise comparisons may be conducted to identify which specific groups have different variances. However,

DRAFT

unlike mean comparisons, there are no widely accepted multiple comparison procedures for variances. Common approaches include:

- Conducting pairwise F-tests with Bonferroni correction: Use adjusted significance level $\alpha^* = \alpha/m$, where $m = k(k-1)/2$ is the number of pairwise comparisons.
- Using confidence intervals for variance ratios with simultaneous confidence level adjustments.
- Examining graphical displays such as side-by-side boxplots or variance plots to identify groups with notably different spreads.

9.7.5 Practical Considerations

When conducting variance tests, several practical issues should be considered:

- **Sensitivity to non-normality:** Variance tests are highly sensitive to violations of the normality assumption. It is advisable to check normality using graphical methods (Q-Q plots, histograms) or formal tests (Shapiro-Wilk, Anderson-Darling) before applying these tests.
- **Sample size:** Small sample sizes can make it difficult to detect departures from normality and reduce the power of variance tests. Larger samples are generally preferred.
- **Robust alternatives:** When normality is questionable, consider using Levene's test, the Brown-Forsythe test, or bootstrap methods, which are more robust to non-normality.
- **One-tailed vs. two-tailed tests:** Choose the appropriate alternative hypothesis based on the research question. Quality control problems often involve one-tailed tests (e.g., testing whether variance exceeds a maximum acceptable level).
- **Multiple comparisons:** When comparing variances across more than two groups, use Bartlett's test (for normal data) or Levene's test (for non-normal data) rather than conducting multiple pairwise F-tests, to control the family-wise error rate.
- **Effect size:** Statistical significance does not always imply practical significance. Consider reporting effect sizes such as variance ratios or coefficients of variation to assess the practical importance of differences.
- **Assumption checking:** Always verify assumptions before conducting tests. For Bartlett's test, normality is critical. For F-tests, both normality and independence should be verified.

DRAFT

9.8 Power of the Test

The power of a statistical test is a key measure of its effectiveness in detecting a true effect when it exists. It is denoted by $1 - \beta$, where β is the probability of committing a Type II error, the power reflects the likelihood of correctly rejecting a false null hypothesis. For example, if a test has a power of 0.80, it means there is an 80% chance of detecting an effect if it is present.

Power of the test: The power of a test is the probability of correctly rejecting the null hypothesis when it is false. That is,

$$\begin{aligned} \text{Power} &= P(\text{reject } H_0 | H_0 \text{ false}) \\ &= 1 - \beta \end{aligned}$$

A higher power indicates a greater likelihood of identifying a true effect.

Increasing the sample size or effect size, or decreasing variability, generally enhances the test's power. However, lowering α to reduce Type I errors can also reduce power, highlighting the need for careful consideration in test design. Ensuring adequate power is essential for reliably detecting meaningful effects and making informed conclusions.

- The power of the two-tailed test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ for the specific alternative $\mu = \mu_1$, where the underlying distribution is normal and the population variance σ^2 is assumed known, is given exactly by

$$\text{Power} = \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \quad (9.27)$$

  and approximately by

$$\text{Power} \approx \Phi\left(-z_{\alpha/2} + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right) \quad (9.28)$$

  where Φ is the cumulative distribution function of the standard normal distribution and z_α is the critical values from the standard normal distribution corresponding to the significance level α.

- The power of the left-tail test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$:

$$\text{Power} = \Phi\left(-z_{\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) \quad (9.29)$$

- The power of the right-tail test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$:

$$\text{Power} = \Phi\left(-z_\alpha + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \tag{9.30}$$

Problem 9.19. *(**Obstetrics**) Compute the power of the test for the birthweight data in Problem 9.3 with an alternative mean of 115 ounces (oz) and $\alpha = 0.05$, assuming the true standard deviation = 24 oz.*

Solution

We have $\mu_0 = 120$ oz, $\mu_1 = 115$ oz, $\alpha = 0.05$, $\sigma = 24$, $n = 100$. Thus,

$$\begin{aligned}\text{Power} &= \Phi\left[-z_{0.05} + (120 - 115)\sqrt{100}/24\right] \\ &= \Phi\left[-1.645 + 5 \times 10/24\right] \\ &= \Phi(0.438) = 0.669\end{aligned}$$

Therefore, there is about a 67% chance of detecting a significant difference using a 5% significance level with this sample size.

Problem 9.20. *(**Cardiovascular Disease, Pediatrics**) Using a 5% level of significance and a sample of size 10, compute the power of the test for the cholesterol data in Problem 9.6, with an alternative mean of 190 mg/dL, a null mean of 175 mg/dL, and a standard deviation of 50 mg/dL.*

Solution

We have $\mu_0 = 175$ oz, $\mu_1 = 190$ oz, $\alpha = 0.05$, $\sigma = 50$, $n = 10$. Thus,

$$\begin{aligned}\text{Power} &= \Phi\left[-z_{0.05} + (190 - 175)\sqrt{10}/50\right] \\ &= \Phi\left[-1.645 + 15 \times \sqrt{10}/50\right] \\ &= \Phi(-0.696) = 0.243\end{aligned}$$

Therefore, the chance of finding a significant difference in this case is only 24.3%.

Problem 9.21. *(**Cardiovascular Disease, Pediatrics**) Compute the power of the test for the cholesterol data in Problem 9.6 with a significance level of $\alpha = 0.01$ vs. an alternative mean of190 mg/dL.*

Solution

We have $\mu_0 = 175$ oz, $\mu_1 = 190$ oz, $\alpha = 0.01$, $\sigma = 50$, $n = 10$. Thus,

$$\begin{aligned}\text{Power} &= \Phi\left[-z_{0.01} + (190 - 175)\sqrt{10}/50\right] \\ &= \Phi\left[-2.326 + 15 \times \sqrt{10}/50\right] \\ &= \Phi(-1.377) = 0.08\end{aligned}$$

DRAFT

which is lower than the power of 24.3% for $\alpha = 0.05$, computed in Problem 9.20. What does this mean? It means that if the α level is lowered from 0.05 to 0.01, the β error will be higher or, equivalently, the power, which decreases from 0.243 to 0.08, will be lower.

Factors Affecting the Power

The power of a test is influenced by several factors, including

(1). **Significance level (α):** If the significance level is made smaller (α decreases), z_α increases and hence the power decreases.

(2). **Effect size (d):** If the alternative mean is shifted farther away from the null mean ($d = |\mu_0 - \mu_1|$ increases), then the power increases.

(3). **Data variability:** If the standard deviation of the distribution of individual observations increases (σ increases), then the power decreases.

(4). **Sample size (n):** If the sample size increases (n increases), then the power increases.

9.9 Sample Size Estimation for the Mean Test

Estimating the appropriate sample size for a mean test is crucial to ensure that the test has sufficient power to detect a significant difference. The sample size needed depends on the desired level of statistical significance, the power of the test, the population standard deviation, and the expected difference in means.

To estimate the required sample size, we need the quantity of the following:

- α: Significance level (commonly 0.05).
- β: Type II error rate (commonly 0.2, which gives a power of 0.8).
- σ: Population standard deviation.
- d: Minimum detectable difference in means (effect size).
- $z_{\alpha/2}$: Critical value from the standard normal distribution for a two-tailed test.
- z_β: Critical value from the standard normal distribution corresponding to the power $(1 - \beta)$ of the test.

DRAFT

9.9.1 When Testing for the Mean of a Normal Distribution (One-Sided Alternative)

Suppose we wish to test

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 \end{cases}$$

where the data are normally distributed with mean μ and known variance σ^2. The sample size needed to conduct a **one-sided test** with significance level α and probability of detecting a significant difference with power $100(1-\beta)\%$ is

$$n = \frac{(z_\beta + z_\alpha)^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

where $d = \mu_0 - \mu_1$.

Problem 9.22. ***(Obstetrics)*** *Consider the birthweight data in Problem 9.3. Suppose that $\mu_0 = 120$ oz, $\mu_1 = 115$ oz, $\sigma = 24$, $\alpha = .05$, $1 - \beta = 0.80$, and we use a one-sided test. Compute the appropriate sample size needed to conduct the test.*

Solution

Since the power $1 - \beta = 0.80$ then $\beta = 0.20$. Therefore, the sample size is

$$n = \frac{(z_\beta + z_\alpha)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(z_{0.20} + z_{0.05})^2 \times 24^2}{(120 - 115)^2} = \frac{(0.84 + 1.645)^2 \times 24^2}{(120 - 115)^2}$$
$$= 142.3 \approx 143$$

The sample size is always **rounded up**, so we can be sure to achieve at least the required level of power (in this case, 80%). Thus, a sample size of 143 is needed to have an 80% chance of detecting a significant difference at the 5% level if the alternative mean is 115 oz and a one-sided test is used.

Problem 9.23. ***(Cardiovascular Disease, Pediatrics)*** *Consider the cholesterol data in Problem 9.6. Suppose the null mean is 175 mg/dL, the alternative mean is 190 mg/dL, the standard deviation is 50, and we wish to conduct a one-sided significance test at the 5% level with a power of 90%. How large should the sample size be?*

Solution

Since the power $1 - \beta = 0.90$ then $\beta = 0.10$. Therefore, the sample size is

$$n = \frac{(z_\beta + z_\alpha)^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(z_{0.10} + z_{0.05})^2 \times 50^2}{(190 - 175)^2} = \frac{(1.28 + 1.645)^2}{15^2} = 95.1$$
$$\approx 96$$

Thus, 96 people are needed to achieve a power of 90% using a 5% significance level.

DRAFT

9.9.2 Sample Size Estimation When Testing for the Mean of a Normal Distribution (Two-Sided Alternative)

Suppose we wish to test

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 \end{cases}$$

where the data are normally distributed with mean μ and known variance σ^2. The sample size needed to conduct a **two-sided test** with significance level α and probability of detecting a significant difference with power $100(1-\beta)\%$ is

$$n = \frac{(z_\beta + z_{\alpha/2})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

Note that this sample size is always larger than the corresponding sample size fora one-sided test, because $z_{\alpha/2}$ is larger than z_α.

Problem 9.24. ***(Cardiology)*** *Consider a study of the effect of a calcium-channel-blocking agent on heart rate for patients with unstable angina. Suppose we want at least 80% power for detecting a significant difference if the effect of the drug is to change mean heart rate by 5 beats per minute over 48 hours in either direction and $\sigma = 10$ beats per minute. How many patients should be enrolled in such a study?*

Solution

We assume $\alpha = 0.05$ and $\sigma = 10$ beats per minute.. We intend to use a two-sided test because we are not sure in what direction the heart rate will change after using the drug. Therefore, the sample size is estimated using the two-sided formulation. We have

$$n = \frac{(z_\beta + z_{\alpha/2})^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(z_{0.20} + z_{0.025})^2 \times 10^2}{5^2} = \frac{(0.84 + 1.96)^2 \times 100}{25}$$
$$= 31.36 \approx 32$$

Thus, 32 patients must be studied to have at least an 80% chance of finding a significant difference using a two-sided test with $\alpha = 0.05$ if the true mean change in heart rate from using the drug is 5 beats per minute.

Single Proportion Test

A **Single Proportion Test** is a statistical hypothesis test used to determine whether the proportion of a population differs from a hypothesized value p_0. This type of test is commonly used in situations where we are interested in the proportion of a certain characteristic in a population.

DRAFT

Hypotheses

- **Two-tailed test:** $H_0 : p = p_0$ vs $H_1 : p \neq p_0$
- **Left-tailed test:** $H_0 : p = p_0$ vs $H_1 : p < p_0$
- **Right-tailed test:** $H_0 : p = p_0$ vs $H_1 : p > p_0$

Assumptions

- The sample is drawn randomly from the population.
- The observations are independent.
- The sample size is large enough such that $np_0 \geq 5$ and $n(1 - p_0) \geq 5$.

Test Statistic

The test statistic is based on the sample proportion $\widehat{p} = \dfrac{x}{n}$, where x is the number of successes and n is the sample size. The test statistic is:

$$Z = \frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

Under H_0, the test statistic Z follows a standard normal distribution $Z \sim N(0, 1)$.

Decision Rule

Let z_{cal} is the observed value of test statistic Z. At significance level α:

- **Two-tailed test:** Reject H_0 if $|z_{\text{cal}}| \geq z_{\alpha/2}$
- **Left-tailed test:** Reject H_0 if $z_{\text{cal}} < -z_\alpha$ or $|z_{\text{cal}}| \geq z_\alpha$
- **Right-tailed test:** Reject H_0 if $z_{\text{cal}} \geq z_\alpha$

where z_α and $z_{\alpha/2}$ are the critical values from the standard normal distribution.

p-value

- **Two-tailed test:** p-value $= 2 \cdot P(Z > |z_{\text{cal}}|)$
- **Left-tailed test:** p-value $= P(Z < z_{\text{cal}})$
- **Right-tailed test:** p-value $= P(Z > z_{\text{cal}})$

Reject H_0 if p-value $\geq \alpha$.

DRAFT

Confidence Interval

A $(1-\alpha)\times 100\%$ confidence interval for the true proportion p is:

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

If the hypothesized value p_0 does not fall within this interval, we reject H_0.

Conclusion

- If H_0 is rejected, there is sufficient evidence at the α level of significance to conclude that the population proportion differs from p_0.
- If H_0 is not rejected, there is insufficient evidence to conclude that the population proportion differs from p_0.

Problem 9.25. *Suppose we want to test if the proportion of defective items in a production process is 5%. We take a sample of 200 items and find that 12 are defective.*

Solution

We wish to test whether the proportion of defective items in the production process differs from the claimed value of 5%. The appropriate hypotheses are:

$$H_0 : p = 0.05 \quad \text{vs} \quad H_1 : p \neq 0.05$$

This is a two-tailed test conducted at the significance level $\alpha = 0.05$.

Before applying the test, we verify the large sample conditions. Since $np_0 = 200 \times 0.05 = 10 \geq 5$ and $n(1-p_0) = 200 \times 0.95 = 190 \geq 5$, the normal approximation is appropriate.

The sample proportion of defective items is computed as:

$$\widehat{p} = \frac{x}{n} = \frac{12}{200} = 0.06$$

Substituting into the test statistic formula, we obtain:

$$z_{\text{cal}} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.06 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{200}}} = \frac{0.01}{0.01541} \approx 0.649$$

For a two-tailed test at $\alpha = 0.05$, the critical values are $z_{0.025} = 1.96$, and the corresponding p-value is:

$$p\text{-value} = 2 \cdot P(Z > 0.649) = 2 \times 0.2582 = 0.5164$$

Since $|z_{\text{cal}}| = 0.649 < 1.96$ and the p-value $0.5164 > 0.05$, we fail to reject H_0. At the 5% level of significance, there is insufficient evidence to conclude that the proportion of defective items differs from 5%. The observed proportion $\widehat{p} = 0.06$ is not significantly different from the hypothesized value $p_0 = 0.05$.

DRAFT

Problem 9.26. *A university administrator claims that 15% of students at a particular university are drug addicted. A health researcher, suspecting that the true proportion is higher, randomly selects 300 students and finds that 54 of them are drug addicted. At the 5% level of significance, is there sufficient evidence to support the researcher's claim?*

Solution:

The researcher suspects that the proportion of drug addicted students exceeds the claimed value of 15%. Therefore, the appropriate hypotheses are:

$$H_0 : p = 0.15 \quad \text{vs} \quad H_1 : p > 0.15$$

This is a right-tailed test conducted at the significance level $\alpha = 0.05$.

Before applying the test, we verify the large sample conditions. Since $np_0 = 300 \times 0.15 = 45 \geq 5$ and $n(1 - p_0) = 300 \times 0.85 = 255 \geq 5$, the normal approximation is appropriate.

The sample proportion of drug addicted students is computed as:

$$\widehat{p} = \frac{x}{n} = \frac{54}{300} = 0.18$$

Substituting into the test statistic formula, we obtain:

$$Z = \frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.18 - 0.15}{\sqrt{\dfrac{0.15 \times 0.85}{300}}} = \frac{0.03}{0.02062} \approx 1.455$$

For a right-tailed test at $\alpha = 0.05$, the critical value is $z_{0.05} = 1.645$, and the corresponding p-value is:

$$p\text{-value} = P(Z > 1.455) = 1 - P(Z \leq 1.455) = 1 - 0.9271 = 0.0729$$

Since $z_{\text{cal}} = 1.455 < 1.645$ and the p-value $0.0729 > 0.05$, we fail to reject H_0. At the 5% level of significance, there is insufficient evidence to support the researcher's claim that the proportion of drug addicted students at the university exceeds 15%. Although the observed proportion $\widehat{p} = 0.18$ is slightly higher than the hypothesized value $p_0 = 0.15$, this difference is not large enough to be statistically significant.

9.10 Sample Size Estimation for Proportion Test

In practice, before conducting a study or survey, it is essential to determine an adequate sample size that ensures the results are statistically reliable. Estimating the sample size in advance helps the researcher to control the margin of error, achieve the desired level of precision, and maintain the required power of the test.

DRAFT

Sample Size for Estimating a Proportion

Suppose we wish to estimate the population proportion p with a specified margin of error ME and a confidence level of $(1-\alpha)\times 100\%$. The margin of error is defined as:

$$\text{ME} = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Solving for the sample size n, we obtain:

$$n = \frac{z_{\alpha/2}^2 \cdot p(1-p)}{\text{ME}^2}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level, p is the estimated or assumed population proportion, ME is the desired margin of error, and n is the required sample size. The resulting value of n should always be rounded **up** to the nearest whole number to ensure the margin of error does not exceed the specified value.

Cases for Estimating p

In practice, the value of p used in the formula may not always be known. Two situations commonly arise.

Case 1: Prior Estimate of p is Available. If a prior study, pilot survey, or historical data provides an estimate of p, this value is substituted directly into the formula:

$$n = \frac{z_{\alpha/2}^2 \cdot \widehat{p}(1-\widehat{p})}{\text{ME}^2}$$

Case 2: No Prior Estimate of p is Available. When no prior information about p is available, we use the conservative estimate $p = 0.5$, since the product $p(1-p)$ is maximized at $p = 0.5$, giving the largest possible sample size and thus the most conservative estimate:

$$n = \frac{z_{\alpha/2}^2 \cdot (0.5)(0.5)}{\text{ME}^2} = \frac{0.25 \cdot z_{\alpha/2}^2}{\text{ME}^2}$$

Sample Size for Hypothesis Testing

When the goal is to conduct a hypothesis test rather than estimate a proportion, the sample size must be large enough to achieve a desired **power** $(1-\beta)$ of the test, where β is the probability of a Type II error. The required sample size is given by:

$$n = \frac{\left(z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_{\beta}\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}$$

where p_0 is the hypothesized proportion under H_0, p_1 is the true proportion under H_1 that the test should detect, $z_{\alpha/2}$ is the critical value corresponding to

DRAFT

the significance level α, and z_β is the critical value corresponding to the desired power $(1-\beta)$.

It is important to note that this formula does not change with the choice of power level. What changes is simply the numerical value substituted for z_β. The value of z_β is obtained from the standard normal distribution such that $P(Z < z_\beta) = 1-\beta$. For instance, to ensure 80% power, we set $1-\beta = 0.80$, which gives $\beta = 0.20$ and therefore $z_\beta = z_{0.20} = 0.842$. The commonly used values of z_β for different power levels are summarized below:

Power $(1-\beta)$	β	z_β
80%	0.20	0.842
85%	0.15	1.036
90%	0.10	1.282
95%	0.05	1.645

Table 9.7: Critical values z_β for commonly used power levels.

A higher desired power requires a larger sample size, as the test must be more sensitive to detect a given difference between p_0 and p_1. Similarly, a smaller margin of difference $|p_1 - p_0|$ also demands a larger sample size, since detecting a subtle difference requires more data. The researcher must therefore balance the practical constraints of data collection against the statistical requirements of the study when choosing an appropriate sample size.

Problem 9.27. *Suppose we want to estimate the sample size needed to test whether the proportion of defective items is different from 5% at a 5% significance level with 80% power, and we expect the true proportion to be 8%.*

Solution

We wish to determine the sample size required to detect a difference between the hypothesized proportion $p_0 = 0.05$ and the expected true proportion $p_1 = 0.08$. The test is two-tailed at significance level $\alpha = 0.05$ with a desired power of $1-\beta = 0.80$.

The required sample size formula for a hypothesis test on a proportion is:

$$n = \frac{\left(z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}\right)^2}{(p_1-p_0)^2}$$

Since $\alpha = 0.05$, we have $z_{\alpha/2} = z_{0.025} = 1.960$, and since the desired power is 80%, we have $z_\beta = z_{0.20} = 0.842$. Substituting all known values into the formula:

$$n = \frac{\left(1.960\sqrt{0.05\times 0.95} + 0.842\sqrt{0.08\times 0.92}\right)^2}{(0.08-0.05)^2} = \frac{0.4297}{0.0009} \approx 477.4$$

DRAFT

Rounding up to the nearest whole number, we obtain $n = 478$. Therefore, a sample of at least **478 items** must be selected in order to detect a difference between the true proportion $p_1 = 0.08$ and the hypothesized proportion $p_0 = 0.05$ at the 5% significance level with 80% power.

Remark 9.10.1. *For a finite sample size, you have to adjust the preliminary sample size. In this case, please see Section* **??** *in Chapter 8.*

9.11 Concluding Remarks

In this chapter, we explored the fundamental concepts of hypothesis testing, a critical tool for decision-making in statistics. We began with the essential steps of formulating hypotheses, establishing significance levels, and understanding test statistics. The introduction of p-values and their interpretation highlighted the nuances of drawing conclusions from statistical data.

We delved into various testing methodologies, including tests for means, variance and proportions, each with specific applications and considerations. The discussions on power analysis and sample size estimation underscore the importance of planning in hypothesis testing to ensure robust and reliable results.

As we conclude, it is clear that hypothesis testing not only enhances our ability to make informed decisions based on data but also provides a structured framework for assessing the validity of claims in diverse fields. Mastery of these concepts will empower you to apply statistical reasoning effectively in your future endeavors.

9.12 Chapter Exercises

1. A company claims that their new battery lasts an average of 500 hours. To test this claim, a sample of 20 batteries is tested, and the sample mean is found to be 490 hours with a sample standard deviation of 15 hours. Test the company's claim at a 5% significance level.

2. A researcher wants to test if the average height of a population of adult males is different from 70 inches. A random sample of 30 men is selected, and the sample mean height is found to be 68.5 inches with a standard deviation of 3 inches. Perform a hypothesis test at the 5% significance level.

   (a) State the null and alternative hypotheses.

   (b) Calculate the test statistic.

   (c) Determine the critical value or p-value.

DRAFT

(d) Make a decision to reject or fail to reject the null hypothesis.

3. In a survey of 200 people, 120 indicated that they prefer coffee over tea. Test whether the proportion of people who prefer coffee is different from 50% at the 1% significance level.

   (a) State the null and alternative hypotheses.
   (b) Calculate the sample proportion.
   (c) Calculate the test statistic.
   (d) Determine the critical value or p-value.
   (e) Make a decision to reject or fail to reject the null hypothesis.

4. A pharmaceutical company is testing a new drug that they believe will lower blood pressure by an average of 5 mmHg. The standard deviation of blood pressure in the population is known to be 10 mmHg. Calculate the power of the test if the sample size is 50 and the significance level is 0.05. Calculate the power of the test.

5. A study is being designed to estimate the mean weight of newborns in a hospital. The standard deviation of birth weights is known to be 1.5 kg. How many newborns need to be included in the sample to estimate the mean weight within 0.2 kg with 95% confidence?

   (a) State the desired confidence level and margin of error.
   (b) Determine the critical value for the confidence level.
   (c) Calculate the required sample size.

6. A political pollster wants to estimate the proportion of voters who support a particular candidate. How large a sample is needed to estimate the true proportion within 3% with 95% confidence?

   (a) State the desired confidence level and margin of error.
   (b) Determine the critical value for the confidence level.
   (c) Use an estimated proportion (e.g., 0.5 if no prior estimate is available) to calculate the required sample size.

7. Benzene is a potential carcinogen, and a chemical company wants to determine if the concentration of benzene in the air is greater than 1 ppm. The following sample data represents the concentration of benzene (in ppm) in the air at various locations within the company:

0.21	1.44	2.54	2.97	0.00
3.91	2.24	2.41	4.50	0.15
0.30	0.36	4.50	5.03	0.00
2.89	4.71	0.85	2.60	1.26

DRAFT

You are tasked with testing whether the concentration of benzene is greater than 1 ppm at a 5% level of significance.

8. Arsenic contamination in groundwater is a major environmental concern. A health agency wants to investigate whether the mean arsenic concentration in drinking water exceeds the recommended safe limit of 0.05 ppm. The following sample data (in ppm) are collected from 20 different wells:

0.04	0.07	0.10	0.06	0.03
0.08	0.09	0.05	0.11	0.02
0.06	0.04	0.12	0.09	0.05
0.07	0.10	0.03	0.08	0.06

Assume that the arsenic concentrations are normally distributed.

   (a) State the null and alternative hypotheses.
   (b) At the 5% level of significance, test whether the mean arsenic concentration exceeds 0.05 ppm.
   (c) Compute the test statistic and the corresponding p-value.
   (d) State your conclusion in the context of the problem.

9. Suppose we want to estimate the sample size needed to test whether the mean of a population is different from a specified value. We set the significance level at 5% ($\alpha = 0.05$), the power at 80% ($\beta = 0.2$), the population standard deviation at 10, and the minimum detectable difference in means at 3.

10. A manufacturer claims that their light bulbs last an average of 1,200 hours. A consumer group tests a sample of 30 bulbs and finds a mean lifetime of 1,150 hours with a standard deviation of 100 hours. Test the manufacturer's claim at the 0.05 significance level.

11. A school claims that the average score of its students on a standardized test is 75. A random sample of 50 students has a mean score of 72 with a standard deviation of 10. Conduct a hypothesis test at the 0.01 significance level to determine if the school's claim is valid.

12. A researcher wants to compare the effectiveness of two different teaching methods. Method A has a sample mean of 82 with a standard deviation of 5 from 25 students. Method B has a sample mean of 78 with a standard deviation of 6 from 30 students. Test the hypothesis that the two methods are equally effective at a significance level of 0.05.

DRAFT

13. A study compares the average daily water consumption of two cities. City X has a mean consumption of 150 liters with a standard deviation of 20 liters based on a sample of 40 households. City Y has a mean consumption of 160 liters with a standard deviation of 25 liters based on a sample of 35 households. Conduct a hypothesis test at the 0.05 level.

14. A dietician wants to test the effectiveness of a new diet plan. She measures the weight of 10 participants before and after the diet. The weights (in kg) are as follows:

    - Before: 85, 78, 90, 82, 88, 80, 76, 85, 87, 90
    - After: 83, 76, 87, 80, 84, 78, 75, 82, 86, 89

    Test whether the diet plan has significantly reduced weight at the 0.05 significance level.

15. A medical researcher is testing the effect of a new medication on blood pressure. He measures the blood pressure of 12 patients before and after treatment. The readings (in mmHg) are as follows:

    - Before: 130, 135, 140, 128, 132, 138, 136, 134, 129, 137, 141, 133
    - After: 125, 130, 135, 127, 130, 132, 128, 129, 126, 134, 138, 131

    Conduct a hypothesis test at the 0.01 significance level.

16. Explain the concepts of Type I and Type II errors in the context of hypothesis testing. Provide examples of each based on the exercises above.

17. A new software is claimed to improve productivity. If you decide to reject the null hypothesis that it does not improve productivity (Type I error) when in fact it does not, what are the consequences of such an error?

18. A researcher is conducting a study to evaluate whether a new drug is effective in lowering blood pressure. The null hypothesis is that the drug has no effect on blood pressure (i.e., the mean change in blood pressure is zero), while the alternative hypothesis is that the drug does lower blood pressure. The researcher expects that the drug will lower blood pressure by an average of 5 mmHg, based on previous studies. The population standard deviation for the change in blood pressure is known to be 12 mmHg. The sample size for the study is 40 patients, and the significance level for the hypothesis test is set to 0.05. Given the information above, calculate the power of the hypothesis test to detect a true mean difference of 5 mmHg.

19. A company wants to test whether a new production process increases the average output of widgets. The company conducts a hypothesis test using a one-sample t-test with a significance level of 0.05. The null hypothesis is

DRAFT

that the mean output is 100 widgets per hour, and the alternative hypothesis is that the mean output is greater than 100 widgets per hour. The population standard deviation is known to be 15 widgets. The company expects that the new process will increase the mean output by 5 widgets per hour, i.e., the true population mean is 105 widgets per hour. Given that the sample size is 25, calculate the power of the test to detect a true mean of 105 widgets per hour.

20. A coffee vending machine is designed to dispense coffee with a standard deviation of no more than 0.5 ounces. A random sample of 20 cups yields a standard deviation of 0.65 ounces. At the $\alpha = 0.05$ level, test whether the machine exceeds the specification.

21. Two methods of teaching statistics are compared. Method A is used with 25 students, resulting in test score variance of 144. Method B is used with 31 students, resulting in test score variance of 81. At the $\alpha = 0.01$ level, test whether the variances are equal.

22. Construct a 95% confidence interval for the population variance in Exercise 1.

23. A quality control engineer claims that the variance in the diameter of bolts produced by Machine 1 is less than the variance for Machine 2. A sample of 10 bolts from Machine 1 has $s_1^2 = 0.025$, and a sample of 12 bolts from Machine 2 has $s_2^2 = 0.050$. Test the claim at $\alpha = 0.05$.

24. Explain why the F-test for equality of variances is more sensitive to violations of normality compared to the two-sample t-test for equality of means. What alternatives would you recommend when the normality assumption is violated?

25. A researcher is comparing the variability of response times across five different experimental conditions. The sample sizes are $n_1 = 10$, $n_2 = 12$, $n_3 = 15$, $n_4 = 11$, and $n_5 = 13$, with sample variances $s_1^2 = 25$, $s_2^2 = 36$, $s_3^2 = 30$, $s_4^2 = 42$, and $s_5^2 = 28$, respectively. Use Bartlett's test at $\alpha = 0.05$ to determine whether the population variances are equal.

26. Four brands of batteries are tested for consistency in their lifetimes. Would you use Bartlett's test or Levene's test if:

    (a) The lifetime distributions appear highly skewed based on Q-Q plots?

    (b) The data appear approximately normal with no outliers?

    (c) You are unsure about the distributional form of the data?

    Justify your answers.

DRAFT

27. In Exercise 6, if the null hypothesis is rejected, describe how you would conduct post-hoc pairwise comparisons to determine which specific variances differ. What adjustment would you make to control the overall Type I error rate?

28. A pharmaceutical company tests three formulations of a drug to ensure consistent dosage. Samples yield: Formulation A ($n = 20$, $s^2 = 4.2$), Formulation B ($n = 25$, $s^2 = 6.8$), Formulation C ($n = 22$, $s^2 = 3.9$). Test for homogeneity of variance at $\alpha = 0.10$ using the appropriate method.

29. Discuss the relationship between testing for homogeneity of variance and the assumptions required for one-way ANOVA. Why is it important to test for equal variances before conducting ANOVA?

DRAFT

Chapter 10

Nonparametric Tests of Hypothesis

DRAFT

10.1 Introduction

In previous chapters, we have studied hypothesis tests for means, variances, and other parameters under the assumption that the data follow a specific probability distribution, typically the normal distribution. These methods, known as **parametric tests**, provide powerful tools for statistical inference when their assumptions are satisfied. However, real-world data often violate these assumptions. Customer satisfaction scores may be heavily skewed, website conversion rates follow discrete distributions, and response times frequently exhibit long tails with extreme outliers. When the normality assumption or other parametric requirements are not met, traditional t-tests and ANOVA can produce unreliable or invalid results.

Nonparametric tests, also known as **distribution-free tests**, provide robust alternatives that make minimal assumptions about the underlying data distribution. Rather than analyzing the raw data values directly, these tests typically work with ranks, signs, or frequencies of observations. This fundamental shift in approach makes them remarkably versatile and applicable across diverse statistical scenarios where parametric methods fail.

> **Nonparametric test:** A statistical test that does not require the data to follow a specific probability distribution. Nonparametric tests typically use ranks, signs, or categorical frequencies rather than the original data values.

Consider a common data analysis challenge: comparing user engagement times across two different mobile app interfaces. Engagement time data are typically right-skewed with extreme outliers representing users who leave the app running in the background. A two-sample t-test requires normally dis-

tributed data and can be heavily influenced by these outliers, potentially leading to misleading conclusions. The Mann–Whitney U test, a nonparametric alternative, ranks all engagement times and compares the distributions based on these ranks, making it robust to the skewness and outliers present in the data.

Similarly, when analyzing customer preferences measured on ordinal scales (such as "poor," "fair," "good," "excellent"), parametric methods that assume interval-level measurement are inappropriate. Nonparametric tests such as the Kruskal–Wallis test naturally accommodate ordinal data and provide valid statistical inference without requiring assumptions about the distances between rating categories.

The advantages of nonparametric tests include their flexibility with different data types, robustness to outliers, applicability to small sample sizes, and validity under minimal assumptions. However, these benefits come at a cost: when parametric assumptions are satisfied, nonparametric tests typically have lower statistical power than their parametric counterparts. This means they are less likely to detect true differences when they exist. Fortunately, this power loss is often modest, and in many practical situations where assumptions are violated, nonparametric tests actually provide more reliable and powerful inference than parametric alternatives applied to inappropriate data.

This chapter presents a comprehensive introduction to nonparametric methods, covering tests for paired and independent samples, multiple group comparisons, and goodness-of-fit assessment. Each test is presented with clear guidance on when to use it, how to compute it, and how to interpret the results in practical contexts.

10.2 Characteristics of Nonparametric Tests

Nonparametric tests possess several key characteristics that distinguish them from parametric methods and determine their appropriate applications.

Distribution-Free Property

The defining characteristic of nonparametric tests is that they do not require the data to follow a specific probability distribution. While parametric tests such as the t-test assume normality, nonparametric tests remain valid regardless of the underlying distribution shape. This property makes them particularly valuable when working with skewed distributions, heavy-tailed distributions, or distributions of unknown form.

It is important to note that "distribution-free" does not mean "assumption-free." Most nonparametric tests still require independent observations, and many assume that the distributions being compared have similar shapes, even if those shapes are not specified. Understanding these remaining assumptions is essential for proper application.

DRAFT

Use of Ranks

Most nonparametric tests transform the original data into ranks before conducting the analysis. In this transformation, the smallest observation receives rank 1, the second smallest receives rank 2, and so on. When observations are tied, they are typically assigned the average of the ranks they would have received.

The use of ranks has several important consequences. First, it eliminates the influence of extreme outliers: whether an outlier is 10 times or 1000 times larger than other observations, it receives the same highest rank. Second, ranking preserves only the ordinal information in the data, discarding information about the actual distances between values. This makes ranking appropriate for ordinal data where only the ordering is meaningful, but it also means that some information is lost when ranking interval or ratio-scale data.

Robustness to Outliers

Because nonparametric tests typically work with ranks or signs rather than the original data values, they are inherently robust to outliers. An extreme outlier that would dramatically affect a sample mean has minimal impact on a rank-based test. This robustness is particularly valuable when analyzing real-world datasets that may contain measurement errors, data entry mistakes, or genuinely extreme observations that are difficult to handle with parametric methods.

Applicability to Different Data Types

Nonparametric tests are particularly well-suited to ordinal data, where observations can be ordered but the distances between categories are not meaningful or not known. Examples include:

- Likert-scale responses (strongly disagree, disagree, neutral, agree, strongly agree)
- Educational levels (high school, bachelor's degree, master's degree, doctorate)
- Performance ratings (poor, satisfactory, good, excellent)
- Disease severity classifications (mild, moderate, severe)

Many nonparametric tests can also be applied to continuous data, particularly when distributional assumptions are questionable. Some tests, such as the chi-square test, are specifically designed for nominal (categorical) data.

DRAFT

Power and Efficiency

When the assumptions of parametric tests are satisfied, nonparametric alternatives generally have lower statistical power. This reduction in power is often measured by the **asymptotic relative efficiency** (ARE), which compares the sample sizes needed by the two tests to achieve the same power. For example, the Mann–Whitney U test has an ARE of approximately 0.955 relative to the two-sample t-test when data are normally distributed. This means the nonparametric test requires about 5% more observations to achieve the same power as the parametric test.

However, when parametric assumptions are violated, nonparametric tests often outperform their parametric counterparts. In the presence of heavy-tailed distributions or outliers, the actual power of a t-test can drop substantially, while the nonparametric alternative maintains its validity and power.

Sample Size Flexibility

Many nonparametric tests can be validly applied to quite small sample sizes, sometimes as small as 5 or 6 observations per group. In contrast, parametric tests often require larger samples (commonly 30 or more) for the Central Limit Theorem to ensure approximate normality of test statistics when the underlying data are non-normal. This makes nonparametric tests particularly useful in pilot studies, expensive experiments, or situations where data collection is limited.

For very small samples, exact p-values can often be computed for nonparametric tests. For larger samples, asymptotic approximations (often based on the normal or chi-square distribution) are typically used.

Computational Considerations

Most nonparametric tests involve relatively simple calculations: ranking observations, counting signs, or computing test statistics from ranks. This computational simplicity contributes to their interpretability and made them particularly popular before the widespread availability of statistical software. Modern software packages include nonparametric tests as standard procedures, making them as easy to compute as their parametric counterparts.

10.3 When to Use Nonparametric Tests

Consider using a nonparametric test when one or more of the following conditions hold:

- The data exhibit substantial departures from normality, such as severe skewness, heavy tails, or multimodality, and transformations do not adequately address these issues.

DRAFT

- The data are measured on an ordinal scale, where rankings are meaningful but distances between values are not.
- The sample size is small and the normality assumption cannot be reliably assessed or justified.
- The data contain outliers that are legitimate observations (not measurement errors) but exert undue influence on parametric tests.
- A robust analysis is desired that does not depend on specific distributional assumptions.
- The equal variance assumption required for parametric tests is violated and cannot be corrected.

Conversely, when parametric assumptions are reasonably satisfied and the data are measured on an interval or ratio scale, parametric tests are generally preferred due to their higher statistical power.

Table 10.1 summarizes common nonparametric tests, their purposes, and their parametric equivalents.

DRAFT

Table 10.1: Common Nonparametric Tests and Their Uses

Test	**Purpose**	**Parametric Alternative**	**Data Type**
Sign Test	Test if median equals a specified value or if median difference equals zero (paired data)	One-sample or paired t-test	Ordinal or continuous
Wilcoxon Signed-Rank Test	Test if median difference equals zero for paired samples	Paired t-test	Ordinal or continuous
Mann–Whitney U Test	Compare distributions of two independent groups	Two-sample t-test	Ordinal or continuous
Kruskal–Wallis Test	Compare distributions of three or more independent groups	One-way ANOVA	Ordinal or continuous
Friedman Test	Compare distributions of three or more related groups (repeated measures or matched blocks)	Repeated-measures ANOVA	Ordinal or continuous
Chi-square Test of Independence	Test association between two categorical variables	—	Nominal
Chi-square Goodness-of-Fit Test	Test if observed frequencies match an expected distribution	—	Nominal or ordinal
Fisher's Exact Test	Test association in 2×2 contingency tables with small expected frequencies	—	Nominal
Kolmogorov–Smirnov Test	Test if a sample comes from a specified distribution or if two	—	Continuous

DRAFT

10.4 The Sign Test

The **Sign Test** is a nonparametric method used to test whether the median of a population equals a specified value, or equivalently, whether the median of paired differences is zero. For paired observations, it assesses whether one member of a pair tends to be systematically larger or smaller than the other. Unlike the paired t-test, it does not assume normality; it relies solely on the *direction* of the differences, discarding their magnitude. This makes it robust to outliers, though at the cost of some statistical power compared to tests that also utilize magnitude information.

10.4.1 Test Statistic

Consider n paired observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. For each pair, compute the difference $d_i = X_i - Y_i$ and record only its sign. Pairs with $d_i = 0$ are excluded, as they carry no directional information. Let n_s denote the number of non-zero differences, and define

$$S_+ = \#\{i : d_i > 0\}, \qquad S_- = \#\{i : d_i < 0\}, \qquad n_s = S_+ + S_-.$$

Under the null hypothesis that the median difference $M_d = 0$, a positive or negative sign is equally likely, so

$$S_+ \sim \text{Binomial}(n_s,\, 0.5).$$

10.4.2 Hypotheses and Decision Rule

The hypotheses concern M_d, the median of the population differences.

Two-tailed test:

$$H_0 : M_d = 0 \qquad \text{vs.} \qquad H_1 : M_d \neq 0.$$

Set $k = \min(S_+, S_-)$. The p-value is

$$p = 2\,P(S \leq k \mid H_0), \qquad S \sim \text{Binomial}(n_s, 0.5).$$

One-tailed tests:

- $H_1 : M_d > 0$ (X tends to exceed Y): use S_+ as the test statistic; p-value $= P(S_+ \geq s_+ \mid H_0)$.
- $H_1 : M_d < 0$ (Y tends to exceed X): use S_- as the test statistic; p-value $= P(S_- \geq s_- \mid H_0)$.

Reject H_0 whenever the p-value $\leq \alpha$.

Problem 10.1. *(Effect of a Diet on Patient Weight) A nutritionist tests whether a new diet affects the weight of 10 patients. Weights (in kg) are recorded before and after two months of the diet, as shown in Table 10.2. Using the Sign Test, determine whether the diet significantly reduced weight at $\alpha = 0.05$.*

DRAFT

Table 10.2: Weights Before and After Diet

Patient	**Before** (X_i)	**After** (Y_i)	**Patient**	**Before** (X_i)	**After** (Y_i)
1	78	75	6	82	80
2	85	83	7	79	78
3	90	92	8	84	86
4	88	85	9	91	90
5	95	95	10	77	74

Solution. Since the aim is to detect a *reduction* in weight, we test whether $M_d > 0$, leading to the one-tailed hypotheses

$$H_0 : M_d \leq 0 \qquad \text{vs.} \qquad H_1 : M_d > 0,$$

where p denotes the probability that $X_i > Y_i$ for a randomly chosen pair.

The differences $d_i = X_i - Y_i$ and their signs are as follows.

Patient	X_i	Y_i	d_i	**Sign**
1	78	75	+3	+
2	85	83	+2	+
3	90	92	−2	−
4	88	85	+3	+
5	95	95	0	(excluded)
6	82	80	+2	+
7	79	78	+1	+
8	84	86	−2	−
9	91	90	+1	+
10	77	74	+3	+

Patient 5 is excluded ($d_5 = 0$), giving $n_s = 9$. Among the remaining pairs, $S_+ = 7$ and $S_- = 2$. Since $H_1 : M_d > 0$, the test statistic is $S_+ = 7$ and the p-value is

$$p = P(S_+ \geq 7 \mid H_0) = \sum_{k=7}^{9} \binom{9}{k} (0.5)^9 = \big(36 + 9 + 1\big) \times \frac{1}{512} = \frac{46}{512} \approx 0.0898.$$

Since $p \approx 0.090 > 0.05$, we fail to reject H_0 at the $\alpha = 0.05$ significance level. There is insufficient statistical evidence to conclude that the diet significantly reduced the patients' weight. □

DRAFT

Large Sample Approximation

For large n, the binomial distribution can be approximated by the normal distribution. The Z-score is:

$$Z = \frac{S_+ - \frac{n}{2}}{\sqrt{n/4}}$$

A continuity correction of ± 0.5 is applied as appropriate. The p-value can then be derived from the standard normal table.

Advantages and Limitations

- **Advantages:** The Sign Test demands minimal assumptions, requiring no specific distribution or symmetry.
- **Limitations:** By using only the sign and ignoring the magnitude of differences, it has lower statistical power compared to tests like the Wilcoxon Signed-Rank Test when their assumptions are met.

Python Implementation

```python
from scipy.stats import binom

# Data
before = [78, 85, 90, 88, 95, 82, 79, 84, 91, 77]
after  = [75, 83, 92, 85, 95, 80, 78, 86, 90, 74]

# Differences
differences = [x - y for x, y in zip(before, after)]
print("Differences (d_i = X_i - Y_i):")
for i, d in enumerate(differences, 1):
    sign = "+" if d > 0 else ("-" if d < 0 else "0 (excluded)")
    print(f"  Patient {i:2d}: d = {d:+.1f}  =>  {sign}")

# Exclude zeros
non_zero = [d for d in differences if d != 0]
n_s = len(non_zero)

S_plus  = sum(1 for d in non_zero if d > 0)
S_minus = sum(1 for d in non_zero if d < 0)

print(f"\nn_s = {n_s}  (after excluding zero differences)")
print(f"S+  = {S_plus}  (positive signs)")
print(f"S-  = {S_minus}  (negative signs)")

# One-tailed test: H1: M_d > 0  => use S+
alpha = 0.05
p_value = 1 - binom.cdf(S_plus - 1, n_s, 0.5)  # P(S+ >= 7)
```

```
28
29 print(f"\nTest: H0: M_d <= 0  vs  H1: M_d > 0  (one-tailed)"
       )
30 print(f"Test statistic: S+ = {S_plus}")
31 print(f"p-value = P(S+ >= {S_plus}) = {p_value:.4f}")
32
33 print(f"\nDecision (alpha = {alpha}):")
34 if p_value <= alpha:
35   print("  Reject H0 - significant evidence that the diet
       reduced weight.")
36 else:
37   print(f"  Fail to reject H0 - insufficient evidence (p = {
       p_value:.4f} > {alpha}).")
```

Listing 10.1: Sign Test for Diet Effect on Weight

When to Use the Sign Test

The sign test is appropriate in the following cases.

(i). The data is continuous but not necessarily normally distributed.

(ii). You want to test hypotheses about the population median.

(iii). Sample sizes are small or the distribution is highly skewed

(iv). You're dealing with paired data (before/after comparisons).

(v). The assumptions for t-tests are violated.

10.4.3 Wilcoxon Signed-Rank Test

The **Wilcoxon Signed-Rank Test** is a nonparametric test for paired data or single samples, assessing if the median differs from zero (or a hypothesized value) Unlike the Sign Test, it accounts for both the direction and magnitude of differences, making it more powerful when the underlying assumptions are satisfied. It is applicable when the data need not be normally distributed, provided the differences between paired observations are continuous and symmetric about the median.

Hypotheses

Let M_d denote the median of the population differences $d_i = X_i - Y_i$. The hypotheses are stated as follows:

- **Two-tailed:** $H_0 : M_d = 0 \quad$ vs. $\quad H_1 : M_d \neq 0$
- **One-tailed (right):** $H_0 : M_d \leq 0 \quad$ vs. $\quad H_1 : M_d > 0$
- **One-tailed (left):** $H_0 : M_d \geq 0 \quad$ vs. $\quad H_1 : M_d < 0$

DRAFT

Test Procedure

1. Compute the differences $d_i = X_i - Y_i$ for each paired observation, and discard any pair for which $d_i = 0$. Let n denote the number of remaining non-zero differences.

2. Rank the absolute values $|d_i|$ from smallest to largest, assigning average ranks in the case of ties.

3. Attach to each rank (R_i) the sign of its corresponding difference, and compute:
$$W_+ = \sum_{\{i\,:\,d_i>0\}} R_i, \qquad W_- = \sum_{\{i\,:\,d_i<0\}} R_i$$

4. The test statistic is then determined according to the alternative hypothesis:
$$W = \begin{cases} \min(W_+,\, W_-) & \text{(two-tailed)} \\ W_+ & \text{(one-tailed, right)} \\ W_- & \text{(one-tailed, left)} \end{cases}$$

Decision Rule

Let W_α denote the critical value obtained from the Wilcoxon Signed-Rank table for a given sample size n and significance level α. The null hypothesis is rejected under the following conditions:

- **Two-tailed:** Reject H_0 if $W \leq W_\alpha$
- **One-tailed (right):** Reject H_0 if $W_+ \geq \dfrac{n(n+1)}{2} - W_\alpha$
- **One-tailed (left):** Reject H_0 if $W_- \leq W_\alpha$

Large Sample Approximation

For $n > 20$, use the normal approximation:

$$Z = \frac{W - \mu_W}{\sigma_W}$$

where $\mu_W = \frac{n(n+1)}{4}$ and $\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

Compare Z with standard normal critical values.

Problem 10.2. *(Training Program for Call Handling Time) A company implemented a new training program for its customer service representatives to improve their call handling time. To assess the program's effectiveness, they randomly selected 9 employees and recorded their average call handling time (in minutes) before the training and one month after completing the training.*

DRAFT

Table 10.3: Critical Values for Wilcoxon Signed-Rank Test

n	**One-Tailed Test**		**Two-Tailed Test**	
	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.10$	$\alpha = 0.05$
5	0	—	0	—
6	2	0	2	0
7	3	2	3	2
8	5	3	5	3
9	8	5	8	5
10	10	8	10	8
11	13	10	13	10
12	17	13	17	13
13	21	17	21	17
14	25	21	25	21
15	30	25	30	25
16	35	29	35	29
17	41	34	41	34
18	47	40	47	40
19	53	46	53	46
20	60	52	60	52

Employee	***Before Training*** (X_i)	***After Training*** (Y_i)
1	*5.2*	*4.8*
2	*6.1*	*5.5*
3	*4.5*	*4.7*
4	*7.0*	*6.2*
5	*5.8*	*5.8*
6	*6.5*	*6.0*
7	*5.0*	*4.9*
8	*7.2*	*7.5*
9	*5.9*	*5.3*

DRAFT

Using the Wilcoxon Signed-Rank Test, determine if the new training program significantly reduced the call handling time at a significance level of $\alpha = 0.05$.

Solution. Since the objective is to determine whether the training program *reduces* call handling time, we test whether the median of the paired differences $d_i = X_i - Y_i$ is positive. This leads to the one-tailed hypotheses:

$$H_0 : M_d \leq 0 \qquad \text{vs.} \qquad H_1 : M_d > 0$$

The differences, their absolute values, and signed ranks are summarized in the following table. Employee 5, whose before and after times are identical ($d_5 = 0$), is excluded from the analysis, leaving $n = 8$ non-zero differences. Since Employees 2 and 9 share the same absolute difference of 0.6, they are assigned the average rank $\frac{6+7}{2} = 6.5$.

Employee	**Before**	**After**	d_i	$\lvert d_i \rvert$	**Rank**	**Sign**	**Signed Rank**
1	5.2	4.8	0.4	0.4	4	+	+4
2	6.1	5.5	0.6	0.6	6.5	+	+6.5
3	4.5	4.7	-0.2	0.2	2	−	−2
4	7.0	6.2	0.8	0.8	8	+	+8
5	5.8	5.8	0.0	0.0	—	—	(Excluded)
6	6.5	6.0	0.5	0.5	5	+	+5
7	5.0	4.9	0.1	0.1	1	+	+1
8	7.2	7.5	-0.3	0.3	3	−	−3
9	5.9	5.3	0.6	0.6	6.5	+	+6.5

Summing the positive and negative ranks separately gives

$$W_+ = 4 + 6.5 + 8 + 5 + 1 + 6.5 = 31, \qquad W_- = 2 + 3 = 5,$$

with the check $W_+ + W_- = 36 = \frac{8\times 9}{2}$ confirming the calculations. Since the alternative hypothesis is $M_d > 0$, the relevant test statistic is $W = W_+ = 31$.

From the Wilcoxon Signed-Rank critical value table at $n = 8$ and $\alpha = 0.05$ (one-tailed), the critical value is $W_\alpha = 5$, and the rejection criterion is

$$W_+ \geq \frac{n(n+1)}{2} - W_\alpha = 36 - 5 = 31.$$

Since $W_+ = 31 \geq 31$, we reject H_0 at the boundary of significance. There is sufficient statistical evidence to conclude that the training program significantly reduced the call handling time, although the result lies exactly at the threshold of significance. □

DRAFT

Python Implementation

```python
import numpy as np
from scipy.stats import wilcoxon

# Data
before_training = [5.2, 6.1, 4.5, 7.0, 5.8, 6.5, 5.0, 7.2, 5.9]
after_training = [4.8, 5.5, 4.7, 6.2, 5.8, 6.0, 4.9, 7.5, 5.3]

# Calculate differences (Before - After)
differences = np.array(before_training) - np.array(after_training)
print("Differences:", differences)

# Hypothesis test
print("\nH0: Training has no effect")
print("H1: Training reduces call handling time (one-tailed)")

# Perform Wilcoxon signed-rank test
statistic, p_value = wilcoxon(differences, alternative='greater')
print(f"Test statistic: {statistic}")
print(f"p-value: {p_value:.4f}")

# Decision
if p_value < 0.05:
  print("Decision: Reject H0")
  print("Conclusion: Training significantly reduces call handling time")
else:
  print("Decision: Fail to reject H0")
  print("Conclusion: No significant evidence training reduces call handling time")
```

10.5 The Mann–Whitney U Test

When comparing **two independent groups** without assuming normality or other parametric distribution requirements, the **Wilcoxon Rank-Sum Test** (developed by Frank Wilcoxon in 1945) and the **Mann–Whitney U Test** (developed independently by Henry Mann and Donald Whitney in 1947) are the most widely used nonparametric alternatives to the **independent samples t-test**. While bearing different names, they are **mathematically equivalent** tests designed for the same purpose and yield identical p-values. The choice of name often depends on historical usage within a particular field or the specific

statistical software output.

Both tests assess whether observations in one sample tend to be systematically larger or smaller than observations in the other sample. They achieve this by ranking all observations from both groups combined and then examining if the sum of ranks (Wilcoxon approach) or the number of pairwise comparisons favoring one group (Mann-Whitney U approach) differs significantly from what would be expected under the null hypothesis of identical distributions.

The Mann–Whitney U test is a nonparametric method used to compare two independent groups when the assumptions required for parametric tests cannot be met. The test works by ranking all observations from both groups together, from smallest to largest, and then evaluating whether the ranks are distributed differently between the two groups. Specifically, it tests the null hypothesis that the two populations have identical distributions against various alternative hypotheses.

By relying on ranks rather than raw data values, the Mann–Whitney U test is robust to outliers and non-normal distributions, making it particularly well-suited for ordinal data or continuous data that violate normality assumptions. This robustness comes at the cost of some statistical power when parametric assumptions are actually met.

Hypotheses and Assumptions

The null hypothesis (H_0) states that the distributions of the two populations are identical, meaning randomly selected values from each group have the same probability distribution.

Let F_1 and F_2 denote the distribution functions of the two populations. The hypotheses are

$$H_0 : F_1(x) = F_2(x) \quad \text{for all } x \qquad \text{vs.} \qquad H_1 : F_1(x) \neq F_2(x) \quad \text{for some } x,$$

with one-tailed variants replacing H_1 by $F_1(x) > F_2(x)$ or $F_1(x) < F_2(x)$ as appropriate. The test assumes that the two samples are independent, the observations are at least ordinal, and the two distributions have the same shape so that a shift in location is the only possible difference under H_1.

Test Statistic and Decision Rule

Let n_1 and n_2 denote the sizes of the two samples. All $N = n_1 + n_2$ observations are combined and ranked from smallest to largest, with average ranks assigned in the case of ties. Let R_1 and R_2 denote the sum of ranks for each group, satisfying

$$R_1 + R_2 = \frac{N(N+1)}{2}.$$

DRAFT

The test statistics are

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2,$$

with the check $U_1 + U_2 = n_1 n_2$. The observed test statistic is $U = \min(U_1, U_2)$ for a two-tailed test, or U_1 and U_2 respectively for one-tailed tests. For small samples, H_0 is rejected if $U \leq U_{\text{critical}}$ obtained from the Mann–Whitney table at the chosen α.

Large Sample Approximation

For larger samples ($\min(n_1, n_2) \geq 20$), employ normal approximation to compute p-value:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

Apply continuity correction:

$$Z = \frac{|U - E[U]| - 0.5}{\sqrt{\text{Var}[U]}}$$

and H_0 is rejected if $|Z| \geq z_{\alpha/2}$. Alternatively, reject H_0 if p-value$\leq \alpha$, where p-values are computed using the normal approximation or exact distribution as appropriate.

For tied observations, adjust the variance:

$$\text{Var}[U] = \frac{n_1 n_2}{12}\left[(n_1+n_2+1) - \frac{\sum_{i=1}^{k} t_i(t_i^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right]$$

where k is the number of tie groups and t_i is the size of the i-th tie group.

Therefore, reject H_0 if p-value$\leq \alpha$, where p-values are computed using the normal approximation or exact distribution as appropriate.

Advantages and Limitations

- **Advantages:**
  - Does not require normal distribution.
  - Handles ordinal data and skewed distributions.
- **Limitations:**
  - Less powerful than the t-test when data are normally distributed.

DRAFT

- Cannot provide confidence intervals for the difference in means or medians directly.

Problem 10.3. *A researcher wants to compare the effectiveness of two teaching methods on student test scores. The following data were collected:*

- *Method A: 78, 82, 85, 88, 92*
- *Method B: 75, 80, 83, 86, 90, 95*

Test at $\alpha = 0.05$ *whether there is a significant difference between the two methods.*

Solution.

Since the objective is to determine whether the two teaching methods produce different test score distributions, without assuming normality, we apply the Mann–Whitney U test. The hypotheses are

H_0 : The distributions of test scores are identical for both methods

H_1 : The distributions of test scores are not identical for both methods

All $N = n_1 + n_2 = 5 + 6 = 11$ observations are combined and ranked from smallest to largest, with group membership recorded for each:

Value	75	78	80	82	83	85	86	88	90	92	95
Rank	1	2	3	4	5	6	7	8	9	10	11
Group	B	A	B	A	B	A	B	A	B	A	B

The rank sums for each group are

$$R_1 = 2 + 4 + 6 + 8 + 10 = 30 \quad \text{(Method A)},$$
$$R_2 = 1 + 3 + 5 + 7 + 9 + 11 = 36 \quad \text{(Method B)}.$$

The corresponding U statistics are

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 30 + 15 - 30 = 15,$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 30 + 21 - 36 = 15,$$

giving $U_1 + U_2 = n_1 n_2 = 30$. The test statistic is

$$U = \min(U_1, U_2) = 15.$$

For $n_1 = 5$, $n_2 = 6$, and $\alpha = 0.05$ (two-tailed), the Mann–Whitney critical value table (Table 10.4) gives $U_{\text{critical}} = 3$. Since $U = 15 > U_{\text{critical}}$, we fail to reject H_0. At the 0.05 significance level, there is insufficient evidence to conclude that the two teaching methods produce different test score distributions. □

DRAFT

Python Implementation

```python
import numpy as np
from scipy import stats
import pandas as pd

# Example data from the teaching methods problem
method_a = [78, 82, 85, 88, 92]
method_b = [75, 80, 83, 86, 90, 95]

print("Mann-Whitney U Test in Python")
print("=" * 40)

# Method 1: Using scipy.stats.mannwhitneyu (most common)
print("\n1. Using scipy.stats.mannwhitneyu:")
print("-" * 35)

# Perform the test
statistic, p_value = stats.mannwhitneyu(method_a, method_b
  , alternative='two-sided')

print(f"Test statistic (U): {statistic}")
print(f"P-value: {p_value:.6f}")
print(f"Sample sizes: n1={len(method_a)}, n2={len(method_b
  )}")

# Interpretation
alpha = 0.05
if p_value < alpha:
print(f"Result: Reject H0 (p < {alpha})")
else:
print(f"Result: Fail to reject H0 (p >= {alpha})")
```

Remark 10.5.1. *Note that the Mann-Whitney U test is statistically equivalent to the Wilcoxon rank sum test (not to be confused with the Wilcoxon* ***signed*** *rank sum test, which is for paired data).*

Problem 10.4. *(Comparing Recovery Times Under Two Treatments) A clinical researcher wishes to compare the recovery times (in days) of patients under two different treatments. Eight patients were randomly assigned to Treatment A and seven to Treatment B, yielding the following recovery times:*

Treatment A	*12*	*15*	*17*	*19*	*21*	*24*	*27*	*30*
Treatment B	*10*	*13*	*16*	*18*	*22*	*25*	*28*	

Using the Wilcoxon Rank-Sum Test, determine whether there is a significant difference in recovery times between the two treatments at $\alpha = 0.05$.

Solution. Since the data cannot be assumed normal and the two groups are independent, we apply the Wilcoxon Rank-Sum Test. The hypotheses are

$$H_0 : F_A(x) = F_B(x) \quad \text{for all } x \qquad \text{vs.} \qquad H_1 : F_A(x) \neq F_B(x) \quad \text{for some } x.$$

All $N = 8 + 7 = 15$ observations are combined and ranked jointly, with group membership recorded for each:

Value	**Rank**	**Group**
10	1	B
12	2	A
13	3	B
15	4	A
16	5	B
17	6	A
18	7	B
19	8	A
21	9	A
22	10	B
24	11	A
25	12	B
27	13	A
28	14	B
30	15	A

The rank sums are

$$R_A = 2{+}4{+}6{+}8{+}9{+}11{+}13{+}15 = 68, \qquad R_B = 1{+}3{+}5{+}7{+}10{+}12{+}14 = 52,$$

The U statistics are

$$U_A = 8 \times 7 + \frac{8 \times 9}{2} - 68 = 56 + 36 - 68 = 24,$$
$$U_B = 8 \times 7 + \frac{7 \times 8}{2} - 52 = 56 + 28 - 52 = 32.$$

The test statistic is $U = \min(24, 32) = 24$. From the Mann–Whitney table with $n_1 = 8$, $n_2 = 7$, and $\alpha = 0.05$ (two-tailed), the critical value is $U_{\text{critical}} = 8$. Since $U = 24 > 8$, we fail to reject H_0. At the 5% significance level, there is insufficient evidence to conclude that the two treatments produce different recovery times. □

DRAFT

Table 10.4: Mann–Whitney U Critical Values ($\alpha = 0.05$, two-tailed)

n_1	n_2																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Remark 10.5.2 (Understanding the Mann–Whitney U Value)**.** *Unlike many parametric tests where large calculated value of test statistics indicate greater differences between groups (e.g., a large t_{cal}-value in a t-test), the Mann–Whitney U test works differently. The test produces two U statistics, and we typically report the smaller value. A* ***statistically significant*** *Mann–Whitney test occurs when the observed U value (it is denoted by U_{cal}) is* ***smaller than the critical value****. The U statistic measures the degree of overlap in rank orders between groups—the more separated the groups are in their rankings, the smaller the U_{cal} becomes. Therefore, when interpreting Mann–Whitney results:* ***a statistically significant*** *U* ***(i.e., $U_{cal} < U_{critical}$) indicates a meaningful difference between groups****, with smaller significant U values corresponding to greater rank separation. Note that the U value alone does not indicate which group has*

DRAFT

higher values.

Problem 10.5. *A researcher wants to compare the effectiveness of two drugs in reducing blood pressure. Ten patients are randomly assigned to either Drug A or Drug B, and their systolic blood pressure reduction (in mmHg) is recorded:*

Drug A	*12*	*9*	*14*	*11*	*10*
Drug B	*8*	*7*	*6*	*9*	*5*

Test at the 5% significance level whether there is a difference in effectiveness between the two drugs.

Solution. Since the normality of the data cannot be assumed and the two groups are independent, we apply the Wilcoxon Rank-Sum Test with hypotheses

$$H_0 : F_A(x) = F_B(x) \quad \text{for all } x \qquad \text{vs.} \qquad H_1 : F_A(x) \neq F_B(x) \quad \text{for some } x.$$

All $N = 5+5 = 10$ observations are combined and ranked jointly. The values 9_A and 9_B are tied at positions 5 and 6, and are each assigned the average rank $\frac{5+6}{2} = 5.5$:

Value	5	6	7	8	9	9	10	11	12	14
Rank	1	2	3	4	5.5	5.5	7	8	9	10
Group	B	B	B	B	B	A	A	A	A	A

The rank sums are

$$R_A = 5.5 + 7 + 8 + 9 + 10 = 39.5, \qquad R_B = 1 + 2 + 3 + 4 + 5.5 = 15.5$$

The U statistics are

$$U_A = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_A = 25 + 15 - 39.5 = 0.5,$$
$$U_B = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_B = 25 + 15 - 15.5 = 24.5$$

The test statistic is $U = \min(U_A, U_B) = 0.5$.

From the Mann–Whitney table (Table 10.4) with $n_1 = n_2 = 5$ and $\alpha = 0.05$ (two-tailed), the critical value is $U_{\text{critical}} = 2$. Since $U = 0.5 \leq 2$, we reject H_0. There is sufficient evidence at the 5% significance level to conclude that the two drugs differ in their effectiveness in reducing blood pressure. □

DRAFT

10.6 The Kruskal–Wallis Test

The Kruskal–Wallis test is a non-parametric statistical method used to compare the medians of three or more independent groups. It serves as the non-parametric counterpart to the one-way ANOVA and is appropriate when the assumptions of ANOVA (such as normality and homogeneity of variances) are not met. The Kruskal–Wallis test evaluates whether the samples originate from the same distribution. Specifically, it tests if at least one group tends to have systematically higher or lower values than the others based on ranked data.

When to Use

Use the Kruskal–Wallis test when:

- You have three or more independent groups.
- The dependent variable is at least ordinal (ranks can be assigned).
- The data are not normally distributed, or variances are unequal.

Assumptions

- The observations are independent.
- The response variable is at least ordinal.
- The distributions in each group have the same shape (not strictly required, but affects interpretation).

Hypotheses

Let k denote the number of groups and H_0, H_a the null and alternative hypotheses:

H_0 : The population distributions are identical across groups.
H_a : At least one group tends to have different (higher or lower) ranks.

Test Statistic

The Kruskal–Wallis test statistic H is calculated as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where:

- N is the total number of observations across all groups.

DRAFT

- n_i is the number of observations in group i.
- R_i is the sum of the ranks for group i.

If there are tied ranks, a correction factor is applied to adjust the H statistic.

Distribution of the Test Statistic

Under the null hypothesis, and assuming all group sizes are reasonably large ($n_i \geq 5$), the test statistic H approximately follows a chi-square distribution with $k-1$ degrees of freedom:

$$H \sim \chi^2_{k-1}$$

Decision Rule

- Compute the H statistic.
- Determine the critical value from the chi-square distribution with $k-1$ degrees of freedom at a given significance level α.
- If H exceeds the critical value, reject the null hypothesis.

Interpretation

A large value of H indicates that at least one group tends to have higher or lower ranks than the others, suggesting a difference in distributions. Unlike the Mann–Whitney U test, where a small test statistic indicates difference, in the Kruskal–Wallis test, a **large** H statistic suggests stronger evidence against the null hypothesis.

Remark 10.6.1. *The Kruskal–Wallis test does not specify which groups differ. If the null hypothesis is rejected, post-hoc tests (such as pairwise Mann–Whitney tests with adjusted p-values) are required to determine which groups differ significantly.*

Problem 10.6. *A researcher wants to compare the effectiveness of three different teaching methods on student test scores. The scores (out of 100) for each method are as follows:*

Method A	***Method B***	***Method C***
82, 85, 88, 90	*75, 78, 79, 83*	*92, 95, 97, 99*

Test at the 5% significance level whether there is a difference in student performance among the three teaching methods using the Kruskal–Wallis test.

DRAFT

Solution: To test whether there is a significant difference in student performance among the three teaching methods at the 5% significance level, we apply the Kruskal–Wallis test. We begin by combining all 12 scores and ranking them from 1 (lowest) to 12 (highest):

Score	Method	Rank
75	B	1
78	B	2
79	B	3
82	A	4
83	B	5
85	A	6
88	A	7
90	A	8
92	C	9
95	C	10
97	C	11
99	C	12

The sum of ranks for each method is $R_A = 4 + 6 + 7 + 8 = 25$, $R_B = 1 + 2 + 3 + 5 = 11$, and $R_C = 9 + 10 + 11 + 12 = 42$. Each group has $n_i = 4$ observations, and the total number of observations is $N = 12$. The Kruskal–Wallis test statistic is

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \\ &= \frac{12}{12 \times 13} \left(\frac{25^2}{4} + \frac{11^2}{4} + \frac{42^2}{4} \right) - 3 \times 13 \\ &= \frac{12}{156} \left(\frac{625 + 121 + 1764}{4} \right) - 39 \\ &= \frac{12}{156} \left(\frac{2510}{4} \right) - 39 \\ &= \frac{12}{156} (627.5) - 39 \\ &= 48.27 - 39 = 9.27 \end{aligned}$$

With $k = 3$ methods, the degrees of freedom are $df = k - 1 = 2$. At significance level $\alpha = 0.05$, the critical value from the chi-square distribution

DRAFT

is $\chi^2_{0.05,2} = 5.99$. Since the calculated test statistic $H = 9.27 > 5.99$, we reject the null hypothesis. There is sufficient evidence to conclude that at least one teaching method results in significantly different student scores. The data suggest that Method C tends to produce higher scores (average rank = $42/4 = 10.5$) compared to Method A (average rank = $25/4 = 6.25$) and Method B (average rank = $11/4 = 2.75$). To identify which specific methods differ significantly, post-hoc pairwise comparisons with appropriate adjustments for multiple testing are recommended.

10.7 The Friedman Test

The Friedman Test is a nonparametric statistical test used to detect differences among three or more related groups. It is considered the nonparametric alternative to the repeated-measures ANOVA when the assumptions of normality are not satisfied.

When to Use the Friedman Test

- The data consist of repeated measurements or matched observations.
- There are three or more treatments or conditions.
- The response variable is at least ordinal.
- The observations within each block are dependent, but blocks are independent of one another.

Hypotheses

Let k be the number of treatments (or conditions) and n the number of blocks (or subjects). The hypotheses are:

- Null Hypothesis (H_0): All treatments have the same distribution (no treatment effect).
- Alternative Hypothesis (H_1): At least one treatment differs in distribution.

Test Statistic

The Friedman test uses rank-based methods. For each subject, the treatments are ranked from lowest to highest. Ties are assigned average ranks. The test statistic is given by:

$$\chi^2_F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

where:

DRAFT

- R_j is the sum of the ranks for treatment j,
- n is the number of blocks (subjects),
- k is the number of treatments.

Under the null hypothesis, χ_F^2 approximately follows a chi-square distribution with $k-1$ degrees of freedom, especially when $n \geq 10$.

Problem 10.7 (Comparing Teaching Methods Using the Friedman Test)**.** *An education researcher is interested in evaluating the effectiveness of three different teaching methods (A, B, and C). To do so, she selects five students and exposes each of them to all three methods in a randomized order. After completing each teaching module, the students are given the same assessment, and their scores are recorded as follows:*

Student	***Method A***	***Method B***	***Method C***
1	*80*	*75*	*85*
2	*88*	*82*	*89*
3	*90*	*85*	*87*
4	*76*	*80*	*79*
5	*84*	*78*	*83*

Test at the 5% significance level whether there is any statistically significant difference in the performance across the three teaching methods.

Solution: We apply the Friedman test since the same students (blocks) are exposed to each of the three teaching methods (treatments). To conduct the test, we first rank the scores within each student from lowest to highest. For Student 1, Method B received the lowest score (75) and is ranked 1, Method A is ranked 2, and Method C (85) is ranked 3. Similarly ranking the scores for all students yields:

Student	**Rank A**	**Rank B**	**Rank C**
1	2	1	3
2	2	1	3
3	3	1	2
4	1	3	2
5	3	1	2

The sum of ranks for each method is $R_A = 2+2+3+1+3 = 11$, $R_B = 1+1+1+3+1 = 7$, and $R_C = 3+3+2+2+2 = 12$. With $n = 5$ blocks and

DRAFT

$k = 3$ treatments, the Friedman test statistic is

$$\begin{aligned}\chi_F^2 &= \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \\ &= \frac{12}{5 \cdot 3 \cdot 4}(11^2 + 7^2 + 12^2) - 3 \cdot 5 \cdot 4 \\ &= \frac{12}{60}(121 + 49 + 144) - 60 \\ &= 0.2(314) - 60 = 62.8 - 60 = 2.8\end{aligned}$$

From the chi-square table with $df = k - 1 = 2$ and $\alpha = 0.05$, the critical value is $\chi^2_{0.05,2} = 5.991$. Since $\chi_F^2 = 2.8 < 5.991$, we fail to reject the null hypothesis. There is insufficient evidence at the 5% level to conclude that there is a difference in effectiveness among the three teaching methods. While Method B appears to have lower average ranks and Method C higher average ranks, this variation could reasonably occur by chance when the methods are equally effective.

The Friedman test is a valuable tool in situations where the assumptions of parametric repeated-measures ANOVA are violated. It is robust, easy to compute, and applicable to a variety of fields including medicine, psychology, and education. While it can indicate the presence of differences, it does not specify where those differences lie. For that, a post-hoc test such as the Nemenyi test may be applied.

10.7.1 The Friedman Test vs. The Kruskal–Wallis Test

The Friedman test is the nonparametric alternative to the repeated-measures ANOVA. It is used when the same subjects (or matched experimental units) are observed under three or more treatments and the normality assumption required for repeated-measures ANOVA is not satisfied.

In contrast, when the groups being compared are independent, the appropriate nonparametric alternative to the one-way ANOVA is the Kruskal–Wallis test.

For example, suppose a researcher wants to compare the effectiveness of three teaching methods (A, B, and C). If the same group of students is taught using all three methods and their test scores are recorded, the samples are related. When the normality assumption is questionable, the Friedman test can be used to determine whether the teaching methods differ significantly.

On the other hand, if three different groups of students are taught using methods A, B, and C respectively, the groups are independent. In that case, the Kruskal–Wallis test would be the appropriate nonparametric method for comparing the teaching methods.

DRAFT

10.8 Chi-Square Test for Goodness-of-Fit

The **Chi-Square Goodness-of-Fit test** is used to assess whether the observed frequencies of a categorical variable differ significantly from those expected under a specified theoretical distribution. Typical applications include determining whether a die is fair, whether consumer preferences are equally distributed across product categories, or more generally, whether observed counts conform to any hypothesized probability model.

Hypotheses and Assumptions

Suppose there are k categories with observed frequencies $O_1, O_2, \ldots, O_k$ and expected frequencies $E_1, E_2, \ldots, E_k$ derived from the hypothesized distribution. The hypotheses are

$$H_0 : \text{The observed distribution fits the expected distribution}$$
$$H_1 : \text{The observed distribution does not fit.}$$

The test assumes that the sample is randomly selected, observations are independent, and each expected frequency satisfies $E_i \geq 5$ to ensure the validity of the chi-square approximation.

Test Statistic and Decision Rule

The test statistic measures the overall discrepancy between observed and expected counts:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

which follows, under H_0, a chi-square distribution with $\text{df} = k - 1$ degrees of freedom. The null hypothesis is rejected at significance level α if

$$\chi^2 \geq \chi^2_{\alpha,\, k-1},$$

or equivalently, if the p-value $\leq \alpha$.

Problem 10.8. *(Candy Color Distribution) A candy manufacturer claims that five candy colors occur equally frequently. A random sample of 100 candies yielded the following observed frequencies:*

Color	*Red*	*Green*	*Blue*	*Yellow*	*Orange*
Observed (O_i)	*18*	*22*	*20*	*19*	*21*

Test whether the colors are uniformly distributed at $\alpha = 0.05$.

DRAFT

Solution. Under the manufacturer's claim, each color is equally likely with probability $\frac{1}{5}$, leading to the hypotheses

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = \tfrac{1}{5} \qquad \text{vs.} \qquad H_1 : \text{at least one } p_i \neq \tfrac{1}{5}.$$

With a total sample of $n = 100$, the expected frequency for each category under H_0 is $E_i = 100 \times \frac{1}{5} = 20$, and all $E_i \geq 5$, so the chi-square approximation is valid. The test statistic is

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(18-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(21-20)^2}{20} \\
&= 0.20 + 0.20 + 0 + 0.05 + 0.05 \\
&= 0.50.
\end{aligned}$$

With $\text{df} = k - 1 = 4$ and $\alpha = 0.05$, the critical value is $\chi^2_{0.05,\,4} = 9.49$. Since $\chi^2 = 0.50 \ll 9.49$, we fail to reject H_0. There is no significant evidence that the candy colors are distributed unequally. □

Software Implementation (Optional)

In Python:

```
from scipy.stats import chisquare
observed = [18, 22, 20, 19, 21]
expected = [20, 20, 20, 20, 20]
chisquare(f_obs=observed, f_exp=expected)
```

Problem 10.9. *(Testing Fairness of a Die) A six-sided die is rolled 120 times, resulting in the following observed counts:*

Face	*1*	*2*	*3*	*4*	*5*	*6*
Observed O_i	*18*	*22*	*20*	*16*	*24*	*20*

Is there sufficient evidence at the 5% significance level to conclude the die is not fair?

Solution: We want to test whether a six-sided die is fair based on 120 rolls with observed counts:

Face	1	2	3	4	5	6
Observed	18	22	20	16	24	20

Hypotheses:

DRAFT

H_0 : The die is fair, so each face has probability $\frac{1}{6}$

H_1 : The die is not fair

Under the null hypothesis, the expected count for each face is:

$$E_i = 120 \times \frac{1}{6} = 20$$

The chi-square test statistic is calculated as

$$\begin{aligned}\chi^2 &= \frac{(18-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(20-20)^2}{20} \\ &\quad + \frac{(16-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(20-20)^2}{20} \\ &= \frac{4}{20} + \frac{4}{20} + 0 + \frac{16}{20} + \frac{16}{20} + 0 \\ &= 0.2 + 0.2 + 0 + 0.8 + 0.8 + 0 \\ &= 2.0\end{aligned}$$

Degrees of freedom:

$$df = 6 - 1 = 5$$

At significance level $\alpha = 0.05$, the critical value from the chi-square table is:

$$\chi^2_{0.05,5} = 11.07$$

Since the calculated statistic 2.0 is less than the critical value 11.07, we **fail to reject** the null hypothesis. Therefore, there is no significant evidence to suggest the die is unfair.

□

Problem 10.10. *(Testing Preference for Smartphone Brands)*
A survey of 150 people asked about their preferred smartphone brand. The observed counts were:

Brand	*A*	*B*	*C*	*D*
Observed	*40*	*50*	*35*	*25*

The company claims market shares are equal across brands. Test this claim at the 1% significance level.

DRAFT

Solution

For the smartphone brand preference:

$$H_0 : p_A = p_B = p_C = p_D = \frac{1}{4}; \quad H_1 : \text{not all } p_i \text{ equal}$$

Expected frequencies:

$$E_i = 150 \times \frac{1}{4} = 37.5$$

Test statistic:

$$\begin{aligned}\chi^2_{\text{cal}} &= \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(40-37.5)^2}{37.5} + \frac{(50-37.5)^2}{37.5} + \frac{(35-37.5)^2}{37.5} + \frac{(25-37.5)^2}{37.5} \\ &= \frac{2.5^2}{37.5} + \frac{12.5^2}{37.5} + \frac{(-2.5)^2}{37.5} + \frac{(-12.5)^2}{37.5} \\ &= 0.167 + 4.167 + 0.167 + 4.167 = 8.668\end{aligned}$$

Degrees of freedom:

$$df = 4 - 1 = 3$$

Critical value at $\alpha = 0.01$:

$$\chi^2_{0.01,3} = 11.34$$

Since $8.668 < 11.34$, we fail to reject H_0. There is no significant evidence to reject the claim of equal market shares. □

Problem 10.11. *A medical researcher collects blood type data from 200 patients. The blood groups are recorded with Rh factor, i.e., A+, A-, B+, B-, AB+, AB-, O+, and O-. The expected population proportions for these groups are:*

Blood Group	*A+*	*A−*	*B+*	*B−*	*AB+*	*AB−*	*O+*	*O−*
Proportion	0.30	0.10	0.15	0.05	0.05	0.02	0.28	0.05

The observed counts in the sample are:

Blood Group	*A+*	*A−*	*B+*	*B−*	*AB+*	*AB−*	*O+*	*O−*
Observed	58	22	30	10	8	3	50	19

Test at the 5% significance level whether the observed distribution of blood groups matches the expected population distribution.

DRAFT

Solution: We test whether the observed blood group distribution matches the expected population proportions:

$$H_0 : \text{Observed distribution matches expected proportions}$$

$$H_1 : \text{Observed distribution differs from expected proportions}$$

Expected frequencies are calculated as:

$$E_{A+} = 200 \times 0.30 = 60, \quad E_{A-} = 200 \times 0.10 = 20, \quad E_{B+} = 200 \times 0.15 = 30,$$
$$E_{B-} = 200 \times 0.05 = 10, \quad E_{AB+} = 200 \times 0.05 = 10, \quad E_{AB-} = 200 \times 0.02 = 4,$$
$$E_{O+} = 200 \times 0.28 = 56, \quad E_{O-} = 200 \times 0.05 = 10$$

Calculate the chi-square test statistic:

$$\begin{aligned}\chi^2_{\text{cal}} &= \frac{(58-60)^2}{60} + \frac{(22-20)^2}{20} + \frac{(30-30)^2}{30} + \frac{(10-10)^2}{10} \\ &\quad + \frac{(8-10)^2}{10} + \frac{(3-4)^2}{4} + \frac{(50-56)^2}{56} + \frac{(19-10)^2}{10} \\ &= \frac{4}{60} + \frac{4}{20} + 0 + 0 + \frac{4}{10} + \frac{1}{4} + \frac{36}{56} + \frac{81}{10} \\ &= 0.067 + 0.20 + 0 + 0 + 0.40 + 0.25 + 0.643 + 8.10 = 9.66\end{aligned}$$

Degrees of freedom:

$$df = 8 - 1 = 7$$

Critical value at $\alpha = 0.05$:

$$\chi^2_{0.05,7} = 14.07$$

Since $9.66 < 14.07$, we fail to reject H_0. There is no significant evidence to conclude that the observed blood group distribution differs from the expected population distribution. □

10.9 The Pearson Chi-Square Test for Categorical Data

The Pearson Chi-Square (χ^2) test is a fundamental non-parametric statistical method used to analyze the association between categorical variables. It assesses whether the observed frequencies in a contingency table differ significantly from the expected frequencies under the assumption of independence.

DRAFT

Observed and Expected Frequencies

- **Observed Frequencies (O_{ij}):** These are the actual counts recorded in each cell of the contingency table.
- **Expected Frequencies (E_{ij}):** These represent the counts expected if the null hypothesis of independence is true, calculated by:

$$E_{ij} = \frac{(\text{Row Total}_i) \times (\text{Column Total}_j)}{\text{Grand Total}}$$

Hypotheses

- H_0: There is no association between the two categorical variables (i.e., they are independent).
- H_1: There is an association between the two categorical variables (i.e., they are not independent).

Test Statistic

The test statistic is computed as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where r and c denote the number of rows and columns, respectively.

Degrees of Freedom

The degrees of freedom for the test are given by:

$$df = (r-1)(c-1).$$

Assumptions and Considerations

- **Independence:** Each observation contributes to only one cell in the contingency table.
- **Sufficient Sample Size:** Expected cell frequencies should be at least 5 in most cells to ensure the validity of the chi-square approximation. A common guideline is that no more than 20% of cells should have expected frequencies less than 5, and no cell should have an expected frequency less than 1.

Problem 10.12 (Ovarian Cancer)**.** *Suppose we want to investigate whether oral contraceptive (OC) use is associated with the 5-year incidence of ovarian cancer between January 1, 2013, and January 1, 2018. Women who were disease-free*

DRAFT

Table 10.5: Five-year Incidence of Ovarian Cancer by Oral Contraceptive Use

	Ovarian Cancer		
OC User	Yes	No	Total
Ever user	10	40	50
Never user	12	88	100
Total	22	128	150

on January 1, 2013, are classified into two groups based on their OC use status at that time: ever users and never users. The following data were collected:

Is there evidence to suggest that the 5-year incidence of ovarian cancer differs between ever users and never users of oral contraceptives?

Solution: To determine whether there is an association between oral contraceptive use and ovarian cancer incidence, we apply the **Pearson Chi-Square Test of Independence** as described previously. To perform Pearson Chi-Square Test, the hypotheses are in the following:

Hypotheses

H_0 : There is no association between OC use and ovarian cancer incidence.

H_1 : There is an association between OC use and ovarian cancer incidence.

Calculate Expected Frequencies
Using the formula

$$E_{ij} = \frac{(\text{Row Total}_i) \times (\text{Column Total}_j)}{\text{Grand Total}},$$

we compute the expected counts for each cell:

$$\begin{aligned} E_{11} &= \frac{50 \times 22}{150} = \frac{1100}{150} \approx 7.33, \\ E_{12} &= \frac{50 \times 128}{150} = \frac{6400}{150} \approx 42.67, \\ E_{21} &= \frac{100 \times 22}{150} = \frac{2200}{150} \approx 14.67, \\ E_{22} &= \frac{100 \times 128}{150} = \frac{12800}{150} \approx 85.33. \end{aligned}$$

Compute the Chi-Square Test Statistic

DRAFT

The Pearson chi-square statistic is then calculated using the formula

$$\chi^2_{\text{cal}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

giving

$$\chi^2_{\text{cal}} = \frac{(10 - 7.33)^2}{7.33} + \frac{(40 - 42.67)^2}{42.67} + \frac{(12 - 14.67)^2}{14.67} + \frac{(88 - 85.33)^2}{85.33} \approx 1.71.$$

Determine Degrees of Freedom

$$df = (r-1)(c-1) = (2-1)(2-1) = 1.$$

Using a chi-square distribution table or calculator, the critical value at the 5% significance level with 1 degree of freedom is 3.84.

Decision

Since

$$\chi^2_{\text{cal}} = 1.71 < 3.84,$$

we fail to reject the null hypothesis. Therefore, we conclude that there is insufficient evidence to conclude that oral contraceptive use is associated with the 5-year incidence of ovarian cancer in this sample.

Problem 10.13 (Breast Cancer and Age at First Birth)**.** *A hypothesis has been proposed that breast cancer in women is influenced by the age at first childbirth. Women with at least one birth were divided into two groups: those whose age at first birth was* ≤ 29 *years, and those whose age at first birth was* ≥ 30 *years. Among women with breast cancer (cases), 683 of 3220 (21.2%) had age at first birth* ≥ 30*. Among women without breast cancer (controls), 1498 of 10,245 (14.6%) had age at first birth* ≥ 30*. The observed data are summarized below:*

Table 10.6: Breast Cancer Status by Age at First Birth

	Age at First Birth		
Status	≥ 30	≤ 29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

Is there evidence of an association between breast cancer status and age at first birth?

DRAFT

Solution:
We test the hypotheses:

H_0 : There is no association between breast cancer status and age at first birth.

H_1 : There is an association between breast cancer status and age at first birth.

Assuming independence, the expected frequencies for each cell are calculated by

$$E_{ij} = \frac{(\text{Row Total}_i)(\text{Column Total}_j)}{\text{Grand Total}},$$

which gives

$$\begin{aligned} E_{11} &= \frac{3220 \times 2181}{13465} \approx 521.87, \\ E_{12} &= \frac{3220 \times 11284}{13465} \approx 2698.13, \\ E_{21} &= \frac{10245 \times 2181}{13465} \approx 1659.13, \\ E_{22} &= \frac{10245 \times 11284}{13465} \approx 8585.87. \end{aligned}$$

The Pearson chi-square test statistic is

$$\begin{aligned} \chi^2_{\text{cal}} &= \frac{(683 - 521.87)^2}{521.87} + \frac{(2537 - 2698.13)^2}{2698.13} + \frac{(1498 - 1659.13)^2}{1659.13} \\ &\quad + \frac{(8747 - 8585.87)^2}{8585.87} \\ &\approx 78.02. \end{aligned}$$

The degrees of freedom are

$$df = (2-1)(2-1) = 1.$$

At the 5% significance level, the critical value from the chi-square distribution table is 3.84. Since

$$\chi^2_{\text{cal}} = 78.02 > 3.84$$

we reject the null hypothesis. This indicates strong evidence of an association between breast cancer status and age at first birth.

10.10 Yates's Continuity Correction for 2×2 Contingency Tables

When conducting a chi-square test on 2×2 contingency tables, the standard Pearson chi-square test statistic is derived under the assumption that the underlying distribution of the test statistic is continuous. However, observed frequencies are discrete counts. This discrepancy can lead to an inflated Type I

DRAFT

error rate (falsely rejecting the null hypothesis) when sample sizes are small.

To address this issue, particularly when any expected cell frequency is less than 5 (but not so small as to necessitate Fisher's Exact Test), some statisticians advocate for the application of **Yates's Continuity Correction**. This correction is particularly relevant when the degrees of freedom (df) for the test is 1, as is always the case for a 2×2 table.

The formula for the chi-square test statistic with Yates's correction is given by:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

where O_{ij} represents the observed frequency in cell (i, j) and E_{ij} represents the expected frequency in cell (i, j).

The subtraction of 0.5 from the absolute difference between observed and expected frequencies prior to squaring has the effect of reducing the calculated χ^2 statistic. Consequently, this reduction leads to a slightly larger (more conservative) p-value. A more conservative p-value means that stronger evidence is required to reject the null hypothesis, thereby mitigating the risk of committing a Type I error, especially in situations with limited sample sizes.

Problem 10.14. *A researcher is investigating whether there is a relationship between a person's preferred method of communication (email vs. phone) and their gender (male vs. female). They collected data from a small random sample of 40 individuals. The observed frequencies are presented in the contingency table below:*

Table 10.7: Observed Frequencies of Communication Method by Gender

	Email	**Phone**	**Total**
Male	10	8	18
Female	6	16	22
Total	16	24	40

*Perform a chi-square test to determine if there is a statistically significant association between gender and preferred method of communication at a 0.05 level of significance. You **must** apply Yates's Continuity Correction. Show all necessary calculations, including expected frequencies, the corrected chi-square statistic, and your final conclusion.*

a) State the null and alternative hypotheses.

b) Calculate the expected frequencies for each cell.

DRAFT

c) Calculate the chi-square test statistic using Yates's Continuity Correction.

d) Determine the degrees of freedom.

e) State your conclusion regarding the association between gender and communication method at the 0.05 significance level. (You may use the critical value for χ^2 with $df = 1$ at $\alpha = 0.05$ which is 3.841, or state the p-value if you have access to a calculator.)

Solution

a) State the null and alternative hypotheses.

- **Null Hypothesis (H_0):** There is no association between gender and preferred method of communication. (Gender and communication method are independent.)
- **Alternative Hypothesis (H_1):** There is an association between gender and preferred method of communication. (Gender and communication method are dependent.)

b) Calculate the expected frequencies for each cell. The formula for the expected frequency (E_{ij}) for each cell is:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

- $E_{\text{Male, Email}} = \frac{18 \times 16}{40} = \frac{288}{40} = 7.2$
- $E_{\text{Male, Phone}} = \frac{18 \times 24}{40} = \frac{432}{40} = 10.8$
- $E_{\text{Female, Email}} = \frac{22 \times 16}{40} = \frac{352}{40} = 8.8$
- $E_{\text{Female, Phone}} = \frac{22 \times 24}{40} = \frac{528}{40} = 13.2$

Expected Frequencies Table:

Table 10.8: Expected Frequencies of Communication Method by Gender

	Email	**Phone**	**Total**
Male	7.2	10.8	18
Female	8.8	13.2	22
Total	16	24	40

DRAFT

c) Calculate the chi-square test statistic using Yates's Continuity Correction. The chi-square test statistic with Yates's correction is given by:

$$\begin{aligned}\chi^2_{\text{Yates}} &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}} \\ &= \frac{|10-7.2|-0.5)^2}{7.2}+\frac{(|8-10.8|-0.5)^2}{10.8}+\frac{(|6-8.8|-0.5)^2}{8.8} \\ &\quad +\frac{(|16-13.2|-0.5)^2}{13.2} \\ &= 0.735+0.490+0.601+0.401=2.227\end{aligned}$$

d) Determine the degrees of freedom. For a 2×2 contingency table, the degrees of freedom are

$$df=(2-1)(2-1)=1.$$

e) State your conclusion regarding the association between gender and communication method at the 0.05 significance level.

- Significance Level (α): 0.05
- Critical Value for χ^2 ($df=1,\alpha=0.05$): 3.84

Decision Rule: If the calculated χ^2_{Yates} value is greater than the critical value, reject the null hypothesis.

Conclusion: Since the calculated χ^2_{Yates} value $2.227 < 3.84$, we **fail to reject** the null hypothesis. That is, at the 0.05 level of significance, there is not sufficient statistically significant evidence to conclude that there is an association between a person's gender and their preferred method of communication.

DRAFT

10.11 Fisher's Exact Test for 2×2 Tables

In categorical data analysis, we are often interested in testing whether two binary variables are associated. For example, we may ask whether a treatment is associated with recovery, or whether a risk factor is associated with disease.

When data are summarized in a 2×2 contingency table, and especially when sample sizes are small or some expected cell counts are below 5, the traditional chi-square test may not be valid. In such cases, **Fisher's Exact Test** provides an exact method of inference without relying on large-sample approximations.

The Setting

Suppose we have a total of N individuals categorized along two binary variables, arranged in the following 2×2 table:

	Group A	Group B	Row Total
Category 1	x	$K - x$	K
Category 2	$n - x$	$(N - K) - (n - x)$	$N - K$
Column Total	n	$N - n$	N

Here:

- N is the total number of individuals.
- K is the total number of individuals in Category 1.
- n is the number of individuals in Group A.
- x is the number of individuals in both Group A and Category 1.

Null Hypothesis

We want to test the null hypothesis that there is no association between the two categorical variables. Formally, this is stated as:

$$H_0 : \text{there is no association between the two categorical variables.}$$

Under H_0, the marginals (i.e., row and column totals) are fixed, and all possible tables with these margins are equally likely under a hypergeometric model.

The Hypergeometric Model and "Test Statistic"

Under the null hypothesis, the count $X = x$ (i.e., the number of Category 1 individuals in Group A) follows the hypergeometric distribution:

$$X \sim \text{Hypergeometric}(N, K, n).$$

The probability of observing exactly x individuals from Category 1 in Group A is:

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}.$$

This is the probability of drawing x successes (Category 1) in n draws (Group A), without replacement, from a population of N individuals where K are considered "successes."

Unlike many other statistical tests that rely on a single computed "test statistic" (e.g., a t-statistic or χ^2-statistic) whose distribution is then consulted,

DRAFT

Fisher's exact test directly uses the **probability of observing the given table (or a more extreme one) under the null hypothesis** as its basis for inference. In essence, the probability $P(X = x)$ itself acts as the "test statistic" for each possible table configuration, guiding the calculation of the p-value.

Two-Tailed p-Value

To compute the two-tailed p-value in Fisher's Exact Test, we sum the probabilities of all tables with the same marginal totals that are as unlikely or less unlikely (i.e., have probability $\leq P(\text{observed})$) than the observed table:

$$p\text{-value} = \sum_{\text{all } x' \text{ such that } P(X=x') \leq P(X=x_{\text{obs}})} P(X = x')$$

This exact approach avoids relying on asymptotic distributions and is particularly suitable for small sample sizes.

When to Use Fisher's Exact Test

- When any expected cell count in a 2×2 table is less than 5.
- When the sample size is small and the chi-square approximation is unreliable.
- When an exact p-value is needed instead of an approximation.

Interpretation

If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude there is evidence of an association between the two variables. Otherwise, we fail to reject the null hypothesis.

Problem 10.15. *A retrospective study was conducted among men aged 50–54 in a certain county who died over a 1-month period. The objective was to examine whether high salt intake is associated with death due to cardiovascular disease (CVD). Approximately equal numbers of men who died from CVD (cases) and from other causes (controls) were included.*

The data collected are shown in the following contingency table:

Cause of Death	***High Salt Diet***	***Low Salt Diet***	***Total***
Non-CVD (Controls)	*2*	*23*	*25*
CVD (Cases)	*5*	*30*	*35*
Total	*7*	*53*	*60*

(i). State the null and alternative hypotheses to test whether there is an association between high salt intake and cardiovascular death.

DRAFT

(ii). Explain why the chi-square test of independence may not be appropriate for analyzing this dataset.

*(iii). Using **Fisher's Exact Test**, compute the p-value for testing the null hypothesis of no association between diet type and cause of death.*

(iv). At the 5% significance level, what is your conclusion?

(v). Interpret your result in the context of the study. What does it say about the association between high salt intake and CVD?

Solution

A retrospective study among deceased men aged 50–54 examined the possible association between high salt intake and death due to cardiovascular disease (CVD). The data are displayed in the following 2×2 contingency table:

Cause of Death	**High Salt Diet**	**Low Salt Diet**	**Total**
Non-CVD (Controls)	2	23	25
CVD (Cases)	5	30	35
Total	7	53	60

(i) **Hypotheses:**

H_0 : There is no association between high salt intake and cause of death.

H_1 : There is an association between high salt intake and cause of death.

(ii) **Why the Chi-Square Test is Inappropriate:**

The chi-square test of independence relies on the assumption that all expected cell counts are at least 5. Let us calculate the expected counts under the null hypothesis:

$$E_{\text{Non-CVD, High Salt}} = \frac{25 \times 7}{60} = 2.92$$
$$E_{\text{CVD, High Salt}} = \frac{35 \times 7}{60} = 4.08$$

Both expected counts in the "High Salt" column are less than 5, violating the condition for the chi-square test. Hence, Fisher's Exact Test is more appropriate for small samples or low expected frequencies.

(iii) **Fisher's Exact Test:**

To test the null hypothesis of no association between high salt intake and cause of death, we use **Fisher's Exact Test**, which is based on

DRAFT

the **hypergeometric distribution**.

Let X be the number of individuals with a high salt diet who died from non-CVD causes (i.e., the number of "successes" in the non-CVD row, given a fixed total of high salt cases). Under the null hypothesis, the total number of individuals with high salt intake is fixed at 7, and the distribution of these individuals between the two rows (CVD and non-CVD) is random.

That is,

$$X \sim \text{Hypergeometric}(N = 60, K = 25, n = 7)$$

Where:

- $N = 60$: total number of individuals in the study,
- $K = 25$: number of non-CVD deaths (i.e., number of "successes" in the population),
- $n = 7$: total number of high salt diet individuals sampled (i.e., the sample size),
- $x = 2$: observed number of high salt diet individuals in the non-CVD group.

Let's consider your 2x2 contingency table:

	High Salt Diet	**Low Salt Diet**	**Total**
Non-CVD	2 (x)	23 ($K - x$)	25 (K)
CVD	5	30	35
	($n - x$)	$((N - K) - (n - x))$	($N - K$)
Total	7 (n)	53 ($N - n$)	60 (N)

The probability of observing a specific 2x2 table with fixed marginal totals, assuming the null hypothesis (no association), is given by the hypergeometric distribution:

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

$$= \frac{\binom{25}{x}\binom{35}{7-x}}{\binom{60}{7}} \quad ; \quad x = 0, 1, \ldots, 7.$$

DRAFT

For the observed value $x = 2$:

$$P(\text{observed}) = P(X = 2) = \frac{\binom{25}{2}\binom{35}{5}}{\binom{60}{7}} \approx 0.2522$$

The Concept of p-value in Fisher's Exact Test

To obtain the p-value for a two-tailed Fisher's Exact Test, we must consider all possible tables that could be formed with the same marginal totals as the observed table. The p-value is the sum of the probabilities of all such tables that are "as extreme or more extreme" than the observed table. "More extreme" refers to tables whose individual probability of occurrence under the null hypothesis is less than or equal to the probability of the observed table.

For your given marginal totals, the possible values for cell 'x' (Non-CVD, Non-Diabetic) can range from 0 to 7. We would need to calculate the probability for each of these possible tables:

- $x = 0$: $\begin{bmatrix} 0 & 25 \\ 7 & 28 \end{bmatrix}$; $P(X = 0) = \frac{\binom{25}{0}\binom{35}{7}}{\binom{60}{7}} \approx 0.017412.$
- $x = 1$: $\begin{bmatrix} 1 & 24 \\ 6 & 29 \end{bmatrix}$; $P(X = 1) = \frac{\binom{25}{1}\binom{35}{6}}{\binom{60}{7}} \approx 0.105071.$
- $x = 2$: $\begin{bmatrix} 2 & 23 \\ 5 & 30 \end{bmatrix}$ (Observed Table)
- $x = 3$: $\begin{bmatrix} 3 & 22 \\ 4 & 31 \end{bmatrix}$; $P(X = 3) = \frac{\binom{25}{3}\binom{35}{4}}{\binom{60}{7}} \approx 0.311822.$
- $x = 4$: $\begin{bmatrix} 4 & 21 \\ 3 & 32 \end{bmatrix}$; $P(X = 4) = \frac{\binom{25}{4}\binom{35}{3}}{\binom{60}{7}} \approx 0.214378.$

DRAFT

- $x = 5$: $\begin{bmatrix} 5 & 20 \\ 2 & 33 \end{bmatrix}$; $P(X=5) = \dfrac{\binom{25}{5}\binom{35}{2}}{\binom{60}{7}} \approx 0.081853.$
- $x = 6$: $\begin{bmatrix} 6 & 19 \\ 1 & 34 \end{bmatrix}$; $P(X=6) = \dfrac{\binom{25}{6}\binom{35}{1}}{\binom{60}{7}} \approx 0.01605.$
- $x = 7$: $\begin{bmatrix} 7 & 18 \\ 0 & 35 \end{bmatrix}$; $P(X=7) = \dfrac{\binom{25}{7}\binom{35}{0}}{\binom{60}{7}} \approx 0.001245.$

The p-value is then the sum of the probabilities of the tables whose probabilities are less than or equal to $P(\text{observed})$. Based on a full calculation, the p-value is

$$\begin{aligned} p\text{-value} &= P(X=0) + P(X=1) + P(X=2) + P(X=4) \\ &\quad + P(X=5) + P(X=6) + P(X=7) \\ &\approx 0.017412 + 0.105071 + 0.2522 + 0.214378 + 0.081853 \\ &\quad + 0.01605 + 0.001245 \approx 0.688209. \end{aligned}$$

(iv) **Conclusion at 5% Significance Level:**

Since the p-value (0.688209) is much greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. There is no statistically significant evidence of an association between high salt intake and cardiovascular death.

(v) **Interpretation:**

The results of this analysis suggest that there is no statistically significant relationship between high salt intake and dying from cardiovascular disease in this sample. However, it is important to note that the sample size is small, and this study alone cannot rule out a potential association. Further research with a larger sample size would be needed to make stronger conclusions.

Python Code to Calculate Probability: The following Python code computes the hypergeometric probability $P(X = x)$ using `scipy.stats.hypergeom.pmf`:

```
from scipy.stats import hypergeom

# Parameters
N = 60     # Total people
K = 25     # Non-CVD deaths (successes in population)
```

DRAFT

```
n = 7      # High salt diet people (sample size)
x = 2      # Observed high salt among Non-CVD

# Compute probability
p = hypergeom.pmf(x, N, K, n)

print(f"P(X = {x}) =", round(p, 6))
```

Listing 10.2: Python code to compute $P(X = 2)$ using hypergeometric distribution

Python Code: Fisher's Exact Test

The following Python code performs Fisher's Exact Test on a 2×2 contingency table:

```
from scipy.stats import fisher_exact

# Construct the 2x2 contingency table
# Format: [ [Non-CVD: High salt, Low salt],
#           [CVD: High salt, Low salt] ]
table = [
[2, 23],  # Non-CVD group
[5, 30]   # CVD group
]

# Perform Fisher's Exact Test (two-sided)
odds_ratio, p_value = fisher_exact(table, alternative='two
  -sided')

print("Odds Ratio:", round(odds_ratio, 4))
print("P-value:", round(p_value, 6))
```

Listing 10.3: Fisher's Exact Test using `scipy.stats.fisher_exact`

Output:

$$\text{Odds Ratio} \approx 0.5217, \quad \text{P-value} = 0.688178$$

When to Consider Yates's Correction vs. Fisher's Exact Test

When analyzing a 2×2 contingency table, the choice between the standard Chi-square test, Yates's continuity correction, and Fisher's Exact Test depends largely on the sample size and the expected cell counts.

- **Pearson's Chi-square Test (without correction):** This test is appropriate when the sample size is sufficiently large and all expected cell frequencies are reasonably large (typically each expected frequency

is at least 5). Under these conditions, the Chi-square approximation to the sampling distribution is reliable.

- **Yates's Continuity Correction:** Yates's correction is applied to the Chi-square test for a 2×2 table when sample sizes are moderate and expected cell counts are somewhat small. The correction adjusts for the fact that the Chi-square distribution is continuous while the data in contingency tables are discrete. It reduces the Chi-square statistic and therefore provides a more conservative test.

- **Fisher's Exact Test:** Fisher's Exact Test is recommended when the sample size is small or when one or more expected cell frequencies are less than 5. This test computes the exact probability of observing the given table (or one more extreme) under the null hypothesis of independence. It is based on the hypergeometric distribution and does not rely on large-sample approximations.

General guideline:

- If all expected cell counts ≥ 5: use the Pearson Chi-square test.
- If expected counts are near 5 in a 2×2 table: consider the Chi-square test with Yates's correction.
- If expected counts < 5 or the sample size is small: use Fisher's Exact Test.

10.12 Kolmogorov–Smirnov Test for Goodness-of-Fit

The Kolmogorov–Smirnov (K–S) test is a nonparametric test used to compare a sample distribution with a reference probability distribution (one-sample K–S test), or to compare two empirical distributions (two-sample K–S test). It is particularly useful when the underlying distribution is unknown or when testing for normality or other continuous distributional assumptions.

Empirical Distribution Function

Given a sample $\{X_1, X_2, \ldots, X_n\}$, the empirical distribution function $F_n(x)$ is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}},$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Let $x_1 < x_2 < \cdots < x_n$ be the ordered data. The empirical cdf is defined as:

$$F_n(x_i) = \frac{i}{n}$$

DRAFT

Purpose and Applications

The K–S test is applicable in the following situations:

- **One-sample K–S test:** To determine whether a sample comes from a specified theoretical distribution (e.g., normal, uniform, exponential).
- **Two-sample K–S test:** To test whether two independent samples come from the same continuous distribution.

10.12.1 One-Sample K–S Test

Hypotheses

Let $F(x)$ be the cumulative distribution function (cdf) of the theoretical distribution, and let $F_n(x)$ be the empirical cumulative distribution function (ecdf) of the sample. The hypotheses for the one-sample K–S test are:

- H_0: The sample follows the specified distribution with cdf $F(x)$.
- H_1: The sample does not follow the distribution with cdf $F(x)$.

Test Statistic

Let $F(x)$ be the cdf of the reference distribution. The K–S test statistic is defined as the maximum absolute difference between the empirical cdf and the theoretical cdf:

$$D = \sup_x |F_n(x) - F(x)|,$$

where,

- $F_n(x)$ is the empirical cdf of the sample,
- $F(x)$ is the hypothesized theoretical cdf.

The critical values depend on the sample size n, and the significance level α. The distribution of D under the null hypothesis is known, and tabulated critical values are available.

Critical Values for the Kolmogorov–Smirnov Test

For the one-sample Kolmogorov–Smirnov (K–S) test, the critical value at significance level α is approximated by:

$$D_\alpha = \frac{c(\alpha)}{\sqrt{n}}$$

where n is the sample size, and $c(\alpha)$ is a constant depending on the chosen significance level.

DRAFT

Table 10.9: Critical Values for the One-Sample K–S Test

Significance Level α	**Constant** $c(\alpha)$
0.10	1.22
0.05	1.36
0.01	1.63

Example: For $n = 25$ and $\alpha = 0.05$,

$$D_{crit} = \frac{1.36}{\sqrt{25}} = \frac{1.36}{5} = 0.272$$

Problem 10.16. ***Testing Normality with the One-Sample Kolmogorov–Smirnov Test***

A data analyst collects the following sample of exam scores from a class of 10 students and wants to determine whether the scores follow a standard normal distribution:

$$\{-1.2, -0.9, -0.3, 0.0, 0.1, 0.2, 0.4, 0.7, 0.9, 1.1\}$$

Test the null hypothesis that the data follows a standard normal distribution using the Kolmogorov–Smirnov test at the 5% significance level.

Solution. We test whether the data follow a standard normal distribution, giving the hypotheses

$$H_0 : F(x) = \Phi(x) \quad \text{for all } x \qquad \text{vs.} \qquad H_1 : F(x) \neq \Phi(x) \quad \text{for some } x,$$

where Φ denotes the standard normal cdf. The sample is already ordered, so the empirical cdf is simply $F_n(x_i) = i/n$ for $i = 1, \ldots, 10$. For each observation, we compute $\Phi(x_i)$ from the standard normal table and record the absolute discrepancy $|F_n(x_i) - \Phi(x_i)|$:

DRAFT

x_i	$F_n(x_i)$	$\Phi(x_i)$	$\lvert F_n(x_i) - \Phi(x_i)\rvert$
-1.2	0.1	0.1151	0.0151
-0.9	0.2	0.1841	0.0159
-0.3	0.3	0.3821	0.0821
0.0	0.4	0.5000	0.1000
0.1	0.5	0.5398	0.0398
0.2	0.6	0.5793	0.0207
0.4	0.7	0.6554	0.0446
0.7	0.8	0.7580	0.0420
0.9	0.9	0.8159	0.0841
1.1	1.0	0.8643	0.1357

The K–S statistic is the largest of these discrepancies,

$$D = \max_{1 \le i \le n} |F_n(x_i) - \Phi(x_i)| = 0.1357,$$

attained at $x_{10} = 1.1$. For $n = 10$ and $\alpha = 0.05$, the critical value is

$$D_{\text{crit}} = \frac{1.36}{\sqrt{10}} \approx 0.430.$$

Since $D = 0.1357 \ll 0.430$, we fail to reject H_0. There is no significant evidence that the data deviate from a standard normal distribution. □

10.12.2 Two-Sample K–S Test

Hypotheses

Let $F_n(x)$ and $G_m(x)$ be the empirical distribution functions of two samples of sizes n and m, respectively.

- H_0: Both samples are drawn from the same continuous distribution.
- H_1: The samples come from different distributions.

Test Statistic

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

The null distribution of $D_{n,m}$ is used to compute the p-value, or it can be compared with a critical value based on n and m.

DRAFT

Two-Sample K–S Test Critical Value

For the two-sample K–S test with independent samples of size n_1 and n_2, the critical value is given by:

$$D_\alpha = c(\alpha) \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

The constants $c(\alpha)$ remain the same as in the one-sample case.

10.12.3 Advantages and Limitations

- The K–S test is sensitive to differences in both location and shape of the empirical cdf.
- It does not require binning of data or normality assumptions.
- It is limited to continuous distributions and is less powerful in detecting deviations in the tails.
- The test is distribution-free under the null hypothesis but may be conservative when parameters are estimated from the data.

10.12.4 Implementation in Software

- In **Python (SciPy)**, use `scipy.stats.kstest()` for the one-sample test and `scipy.stats.ks_2samp()` for the two-sample test.

The Kolmogorov–Smirnov test is a flexible and widely used method for assessing the goodness-of-fit of a sample to a theoretical distribution or for comparing two empirical distributions. It plays a fundamental role in exploratory data analysis and model diagnostics in data science.

10.13 Chapter Exercises

1. Briefly explain why nonparametric tests are useful in statistical analysis. Provide at least two scenarios where they are preferred over parametric tests.

2. A mobile app developer implemented a new notification system and measured user response times (in seconds) before and after the change for 10 users:

User	1	2	3	4	5	6	7	8	9	10
Before	2.3	3.1	1.7	2.8	3.4	2.1	2.9	1.9	3.2	2.6
After	1.8	2.9	1.9	2.2	3.1	1.6	2.7	1.8	2.8	2.4

DRAFT

(i) Using the **Sign test**, test at the $\alpha = 0.05$ level whether the new system significantly reduces response times.

(ii) State your conclusion in the context of this study.

3. Given the following pre-test and post-test scores of 10 students, perform the Wilcoxon Signed-Rank Test at 5% level of significance.

Pre: 45, 48, 50, 47, 49, 52, 53, 44, 46, 50
Post: 47, 50, 52, 50, 51, 54, 55, 46, 48, 52

4. **(Mann–Whitney U Test)** Two data science teams used different machine learning algorithms to predict customer churn. Their accuracy scores (%) are:

   Team A (Algorithm 1): 78, 82, 75, 88, 79, 84, 77 Team B (Algorithm 2): 85, 91, 87, 82, 89, 93, 86, 90

   (a) Rank all observations from lowest to highest.

   (b) Calculate the rank sums R_1 and R_2 for each team.

   (c) Compute both U_1 and U_2 statistics.

   (d) Using $\alpha = 0.05$ and the critical value table, determine if there's a significant difference between the algorithms.

   (e) What practical conclusion can you draw about the two algorithms?

5. Two fertilizers are tested on different fields to compare yield. Use the Mann–Whitney U test to determine whether the yields differ significantly between the two fertilizers.

Fertilizer A: 20, 22, 19, 25, 24
Fertilizer B: 28, 30, 27, 29, 31

6. Perform the Kruskal–Wallis Test for the following data on crop yields from three regions:

Region 1: 25, 30, 28
Region 2: 35, 33, 37
Region 3: 40, 38, 42

7. A psychologist evaluates three therapies on anxiety reduction across 6 patients. Use the Friedman Test to assess if there's a significant difference in therapy effectiveness.

8. A die is rolled 60 times with observed frequencies: 12, 8, 10, 11, 9, 10. Use the chi-square test to determine if the die is fair.

DRAFT

9. An online retailer wants to test if there's an association between customer age group and preferred shopping method:

	Mobile App	Website	In-Store
18-30 years	45	25	10
31-50 years	30	40	30
51+ years	15	35	50

   (a) Calculate the expected frequencies for each cell.
   (b) Compute the chi-square test statistic.
   (c) Using $\alpha = 0.05$ and appropriate degrees of freedom, make your statistical decision.
   (d) Interpret the result in business terms.

10. The following contingency table shows student preference for three learning modes across two departments. Test if preference is independent of department.

Mode	Dept A	Dept B	Total
Online	30	20	50
Offline	40	60	100
Hybrid	30	20	50
Total	100	100	200

11. A 2×2 table shows observed frequencies for treatment success/failure. Apply Yates's correction to the chi-square test.

	Success	Failure
Group A	30	20
Group B	20	30

12. A market researcher is interested in whether there's a relationship between a person's age group (under 30 vs. 30 and over) and their preference for a particular smartphone operating system (OS) (OS A vs. OS B). A random sample of 50 individuals was surveyed, and the following data were collected:

Age Group	**Prefers OS A**	**Prefers OS B**	**Total**
Under 30	18	7	25
30 and Over	10	15	25
Total	28	22	50

DRAFT

Is there a statistically significant association between age group and smartphone OS preference at a 0.05 level of significance? Given the sample size and the nature of the data, consider applying Yates's Continuity Correction if appropriate.

13. A small sample study yields the following table. Use Fisher's Exact Test to determine association.

	Disease	No Disease
Exposed	4	1
Not Exposed	1	4

14. Explain how the one-sample Kolmogorov–Smirnov test checks for goodness-of-fit. When is it preferred over the chi-square test?

15. Perform a one-sample K-S test to determine if the sample

$$\{5.1, 5.5, 5.7, 5.9, 6.0, 6.3, 6.8\}$$

follows a normal distribution.

16. Compare the two samples using the two-sample Kolmogorov–Smirnov test:

Sample A: 2.1, 2.4, 2.7, 2.9, 3.0
Sample B: 3.2, 3.4, 3.5, 3.8, 4.0

17. Is there any *association* between smoking habit and Myocardial Infarction (MI)? Use the following data:

	Myocardial Infarction		**Total**
Smoking	Heart attack	No heart attack	Total
Smoker	325	175	500
Non-smoker	175	325	500
Total	500	500	1000

(i). State the null and alternative hypotheses for this study.

(ii). Perform the appropriate statistical test to determine whether there is an association between smoking and myocardial infarction. Show all steps clearly.

(iii). At a 5% significance level, what is your conclusion?

(iv). Interpret the result in the context of the study.

DRAFT

Chapter 11

Correlation and Regression Analysis

DRAFT

11.1 Introduction

In the realm of data science, understanding the relationships between variables is crucial for deriving actionable insights from data. This chapter provides a comprehensive overview of **correlation** and **regression analysis**, fundamental techniques in statistical modeling that are essential for data-driven decision-making. Correlation analysis allows data scientists to quantify the **strength** and **direction** of (linear) relationships between two variables, offering a preliminary understanding of their **interdependencies**.

We begin by exploring scatter diagrams, which offer a visual representation of relationships between variables, followed by discussions on covariance and correlation coefficients that formalize these relationships into quantifiable metrics. These foundational concepts pave the way for more advanced regression techniques.

Regression analysis is a powerful tool used to model and predict the behavior of a **dependent variable** based on one or more **independent variables**. This chapter covers simple linear regression as well as multiple linear regression, providing a detailed examination of the assumptions, estimation procedures, and interpretation of regression coefficients. We also delve into evaluating model performance through metrics such as R^2 and adjusted R^2, which are critical for assessing the accuracy and reliability of the model.

To bridge theory and practice, Python code examples are integrated throughout the chapter, demonstrating how to implement these techniques in real-world data science applications. These practical illustrations enhance the understand-

ing of both theoretical concepts and their computational applications, equipping data scientists with the tools needed to effectively analyze and interpret complex datasets.

11.2 Scatter Diagram

A **scatter diagram,** also known as a **scatter plot** or **scatter graph**, is a type of plot that uses Cartesian coordinates to display the values for two different variables. Each data point on the graph represents a pair of observations, with the value of one variable determining the position on the horizontal (x-axis) and the value of the other variable determining the position on the vertical (y-axis).

Typically, the independent variable (the predictor) is placed on the x-axis, and the dependent variable (the one being estimated or predicted) is placed on the y-axis. By plotting these data points, the scatter diagram visually represents the **relationship** between the two variables.

Let's consider an example with data on the height (in cm) and weight (in kg) of a group of individuals. The height and weight pairs might look like this:

Table 11.1: Height and Weight of Individuals

Height (cm)	**Weight (kg)**
150	55
160	62
170	64
180	71
190	73
185	81

The scatter diagram for the data in Table 11.1 is shown in Figure 11.1.

DRAFT

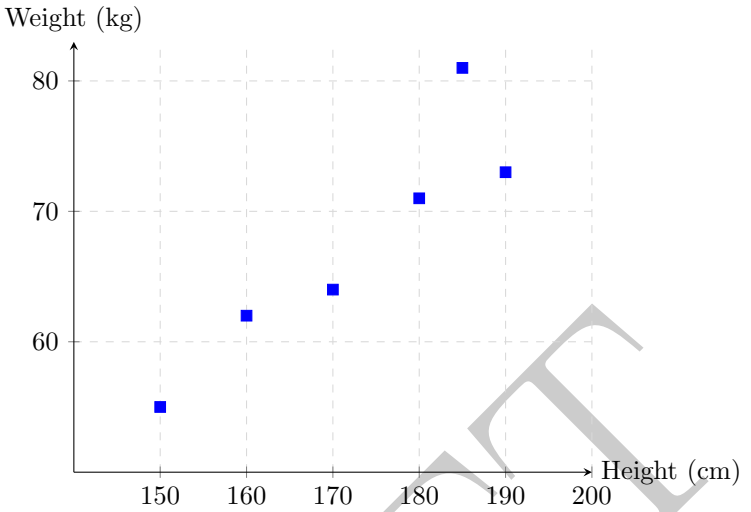


Figure 11.1: Scatter diagram of Height vs. Weight.

Interpreting a scatter plot involves analyzing four key aspects of the data's relationship:

1. Direction of the Relationship

- **Positive Relationship:** If the points trend upwards from left to right, it indicates a positive relationship, meaning that as one variable increases, the other tends to increase as well.
- **Negative Relationship:** If the points trend downwards from left to right, it indicates a negative relationship, meaning that as one variable increases, the other tends to decrease.
- **No Relationship:** If the points are scattered randomly with no clear pattern, there is no discernible relationship between the variables.

2. Strength of the Relationship

- **Strong Relationship:** Points that are tightly clustered around a line or curve suggest a strong relationship between the variables.
- **Weak Relationship:** Points that are more widely spread out but still follow a general trend indicate a weak relationship.
- **No Relationship:** Widely scattered points without a discernible pattern suggest no relationship.

DRAFT

3. **Form of the Relationship**

   - **Linear Relationship:** If the points form a straight line, the relationship is linear. This suggests that a change in one variable is associated with a proportional change in the other.
   - **Non-Linear Relationship:** If the points follow a curved pattern, the relationship is non-linear (e.g., quadratic, exponential).

4. **Outliers**

   - **Outliers:** Points that lie far from the general pattern of the data are called outliers. They may indicate special cases or errors in data collection and can significantly influence the interpretation of the scatter plot.

Interpretation of Figure 11.1

Consider a scatter plot where height (cm) is plotted against weight (kg):

- **Direction:** If the points generally trend upwards, there is a positive correlation between height and weight.
- **Strength:** If the points are closely packed around a line, the correlation is strong; if they are more spread out, the correlation is weaker.
- **Form:** If the points lie on or near a straight line, the relationship is linear.
- **Outliers:** If a point is far away from others (e.g., a height of 190 cm but a weight of 55 kg), it may be an outlier, which requires further investigation.

DRAFT

11.3 Python Code: Scatter diagram

```
import matplotlib.pyplot as plt

# Data
heights = [150, 160, 170, 180, 190, 185]  # Heights in cm
weights = [55, 60, 65, 70, 75, 81]        # Weights in kg

# Create a scatter plot
plt.scatter(heights, weights, color='blue', marker='o')

# Add title and labels
plt.title('Height vs Weight')
plt.xlabel('Height (cm)')
```

```
plt.ylabel('Weight (kg)')

# Show grid
plt.grid(True)

# Display the plot
plt.show()


```

11.4 Correlation Analysis

Correlation analysis is a group of techniques to measure **the strength and direction of the relationship** between two variables. There are different types of correlation coefficients that can be used depending on the nature of the variables being analyzed and the assumptions of the data. Some of the common types of correlation coefficients include:

1. **Covariance:** Measures the direction and magnitude of the linear relationship between two continuous variables.
2. **Pearson correlation coefficient**: Measures **linear relationship** between two continuous variables.
3. **Spearman rank correlation coefficient**: Measures association between ranked variables.
4. **Kendall tau correlation coefficient**: Measures similarity of orderings of data pairs.
5. **Point-biserial correlation coefficient**: Measures association between continuous and binary variables.
6. **Phi coefficient**: Measures association between two binary variables.
7. **Cramér's V**: Measures association between two nominal variables.

11.5 Covariance

A scatter diagram is a powerful tool for visualizing the relationship between two variables. However, it has the following limitations:

- **No Quantitative Measure:** While scatter plots can visually show trends, they do not provide a quantitative measure of the strength or direction of the relationship between the variables. For this, statistical measures like **covariance** or **correlation coefficients** are needed.

DRAFT

- **Subjectivity in Interpretation:** The interpretation of scatter plots can sometimes be subjective, especially in cases of weak or non-linear relationships where the trend is not immediately obvious.
- **Doesn't Reveal Causation:** A scatter plot can show that two variables are related, but it cannot prove that one variable causes the other.

Some of the limitations can be minimized by another popular statistical technique, called covariance.

Covariance: Covariance is a statistical measure that quantifies the degree to which two random variables change together. It assesses whether an increase in one variable corresponds to an increase or decrease in another variable. The sample covariance, denoted by s_{xy}, can be calculated using the formula:

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

where:

- x_i and y_i are the individual data points of the variables x and y,
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively,
- n is the number of data points.

Since,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

then, the covariance is also calculated by using formula

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

Interpretation of the value of Covariance:

- A positive covariance indicates that the variables tend to move in the same direction, meaning that when one variable increases, the other tends to increase as well.
- Conversely, a negative covariance suggests that the variables move in opposite directions—when one variable increases, the other tends to decrease.
- Zero (or close to zero) indicates that there is no (or weak) relationship between two variables.

Problem 11.1. *Consider a dataset containing the number of commercials aired and the corresponding sales volume for a product over ten weeks:*

Table 11.2: Sample Data for the San Francisco Electronics Store

Week	Number of commercials (X)	Sales Volume ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

To further analyze this data, follow these steps:

(i). Draw a scatter diagram of the data points representing the relationship between the number of commercial advertisements and the sales volume.

(ii). Based on the scatter diagram, describe the observed trend or pattern in the data.

(iii). Calculate the sample covariance between the number of commercial advertisements and the sales volume to quantitatively assess the degree of association between these two variables.

(iv). Interpret the sample covariance value in terms of the strength and direction of the relationship between the number of commercial advertisements and the sales volume.

DRAFT

Solution

(i). The scatter plot of Number of commercials (X) and Sales Volume ($100s) is presented in Figure 11.2.

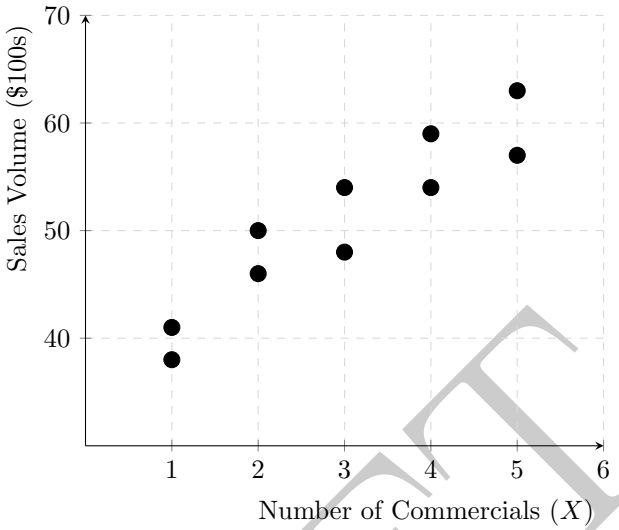


Figure 11.2: Scatter Plot of Number of Commercials vs. Sales Volume

(ii). The scatter plot in Figure 11.5 illustrates the relationship between the number of commercials aired and the sales volume:

- **Positive Relationship**: The plot shows that as the number of commercials increases, the sales volume generally increases as well, indicating a positive relationship between the two variables.
- **Data Points Distribution**: The data points form an upward slope, suggesting that higher numbers of commercials are associated with higher sales volumes. The points are not perfectly aligned, indicating some variability in the relationship.
- **Strength of Relationship**: Although the trend is positive, the relationship is not perfectly linear, suggesting that other factors might influence the sales volume or that the relationship may not be strictly linear.
- **Outliers**: There are no significant outliers, indicating a relatively stable relationship between the two variables across the observed data.

(iii).
Let's calculate the sample covariance for the given data in Table 11.3:

DRAFT

Table 11.3: Data Table

i	x_i	y_i	x_iy_i
1	2	50	100
2	5	57	285
3	1	41	41
4	3	54	162
5	4	54	216
6	1	38	38
7	5	63	315
8	3	48	144
9	4	59	236
10	2	46	92
Total	30	510	1629

$$\text{Mean of } X: \quad \bar{x} = \frac{30}{10} = 3$$
$$\text{Mean of } Y: \quad \bar{y} = \frac{510}{10} = 51$$

Therefore,

$$s_{xy} = \frac{1629 - 10 \times 3 \times 51}{10 - 1} = \frac{99}{9} = 11$$

(iv). The positive covariance of 11 suggests a positive relationship between the number of commercials aired and the sales volume, indicating that as the number of commercials increases, the sales volume tends to increase as well.

□

While covariance measures the direction of the linear relationship between two variables, it has a significant limitation: its value depends on the units of measurement of the variables. For example, if we measure income in dollars versus thousands of dollars, the covariance will change dramatically, making it difficult to interpret the strength of the relationship or compare covariances across different pairs of variables. A covariance of 1000 might indicate a strong relationship for one pair of variables but a weak relationship for another, depending on their scales. This lack of standardization makes covariance an un-

DRAFT

reliable measure for assessing the strength of association. To address this limitation, we use the correlation coefficient, which will be discussed in the next section.

11.6 Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient is a statistical measure that indicates the extent to which two variables fluctuate together. In many contexts, especially when dealing with linear relationships between continuous variables, Pearson's Correlation Coefficient is used and often referred to simply as the **correlation coefficient**.

Pearson's Correlation Coefficient: Pearson's Correlation Coefficient is a measure of the strength of the **linear relationship** between two continuous variables. It is denoted by r or r_{xy} and defined as

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. \tag{11.1}$$

It can be easily shown that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y};$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2;$$

and

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

Therefore, the formula given in Equation (11.1) can be written as

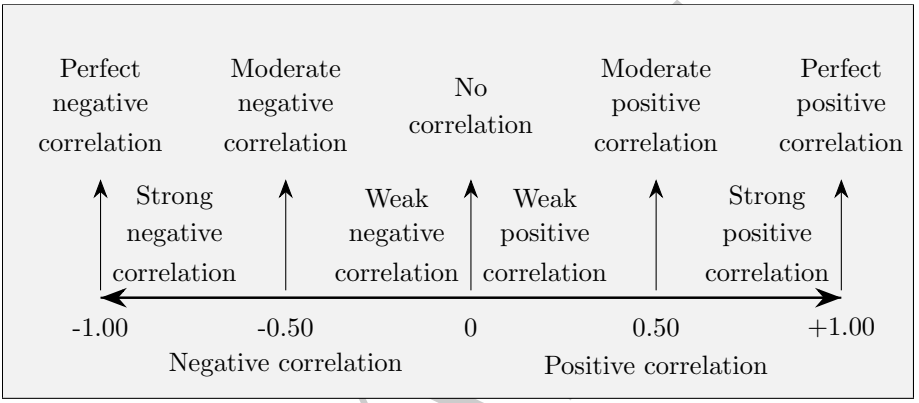
$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \\ &= \frac{s_{xy}}{s_x s_y} \end{aligned}$$

DRAFT

where $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y.

11.6.1 Interpretation of the value of Correlation Coefficient

The following drawing summarizes the strength and direction of the correlation coefficient



Problem 11.2. *Consider Problem 11.1. Calculate the correlation coefficient and interpret the results.*

Solution

From the solution of the Problem 11.1, we have the sample covariance $s_{xy} = 11$. The calculations for the variance are presented in Table 11.4.

We can compute the sample standard deviations for the two variables:

$$s_x = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \sqrt{\frac{1}{9}\left(110 - 10 \times 3^2\right)} = 1.49$$

$$s_y = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = \sqrt{\frac{1}{9}\left(26576 - 30 \times 51^2\right)} = 7.93$$

Hence, the sample correlation coefficient equal

$$r = \frac{s_{xy}}{s_x s_y} = \frac{11}{1.49 \times 7.93} = 0.93$$

which indicates a strong positive linear relationship between number of commercials and sales volume.

DRAFT

Table 11.4: Data Table

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	50	4	2500	100
2	5	57	25	3249	285
3	1	41	1	1681	41
4	3	54	9	2916	162
5	4	54	16	2916	216
6	1	38	1	1444	38
7	5	63	25	3969	315
8	3	48	9	2304	144
9	4	59	16	3481	236
10	2	46	4	2116	92
	30	510	110	26576	1629

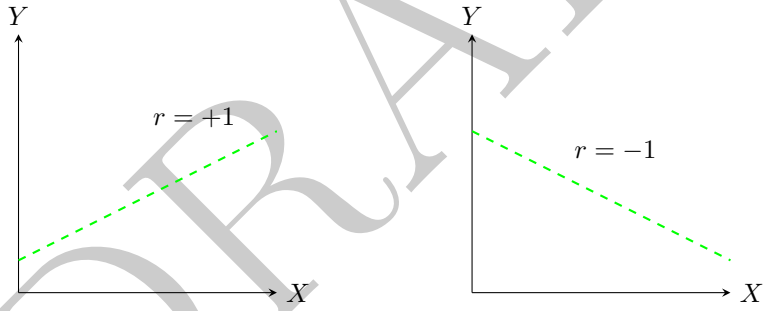


Figure 11.3: Perfect Positive ($r = +1$) and Perfect Negative $r = -1$ Linear Relationship.

11.6.2 Properties of the Correlation Coefficient

The properties of Pearson's correlation coefficient include:

1. **Range:** The correlation coefficient r lies in the range $-1 \leq r \leq 1$. See the Panel A in Figure 11.3.
    - $r = 1$: Indicates a perfect positive linear correlation between variables. See the Panel B in Figure 11.3.

DRAFT

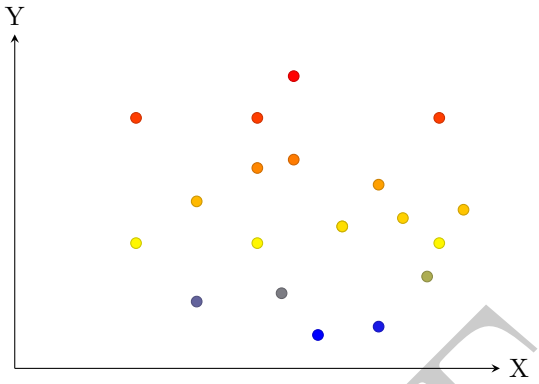


Figure 11.4: No or weak Relationship $r \to 0$

- $r = -1$: Indicates a perfect negative linear correlation between variables.
- $r = 0$: Indicates no linear correlation between variables. See Figure 11.4.
- $|r|$ close to 1 indicates a strong relationship.
- $|r|$ close to 0 indicates a weak relationship.

2. **Symmetry:** The correlation between X and Y is the same as between Y and X:
$$r(X, Y) = r(Y, X)$$

3. **Unit-Free:** The correlation coefficient is a dimensionless number, meaning it does not depend on the units of the variables.

4. **Unaffected by Change of Origin and Scale:** The correlation coefficient remains unchanged if we add a constant to one or both variables, or if we multiply one or both variables by a positive constant:
$$r(aX + b, cY + d) = r(X, Y)$$
where a, b, c, and d are constants, with $a > 0$ and $c > 0$.

5. **Linear Relationship:** The correlation coefficient measures the strength and direction of a linear relationship between two variables. It does not capture non-linear relationships.

6. **Sensitivity to Outliers:** The correlation coefficient is **not robust** to outliers. A single outlier can significantly distort the value of r and provide a misleading interpretation of the relationship between the variables.

DRAFT

7. **Dependence on the Type of Data:** The correlation coefficient is appropriate for continuous data but may not be suitable for ordinal or categorical data without modification.

8. **Pairwise Comparisons:** The correlation coefficient is computed for pairs of variables and does not extend naturally to more than two variables.

Problem 11.3. *The following sample of observations were randomly selected.*

x	4	5	3	6	10
y	4	6	5	7	7

(a). Draw a scatter diagram. Comment on your output.

(b). Determine the correlation coefficient and interpret the relationship between x and y. Interpret this statistical measure.

Solution

(a)
The scatter plot given in Figure 11.5, visualizes the relationship between two variables, X and Y, for the given sample of observations. Each point on the scatter plot represents a pair of x and y values.

DRAFT

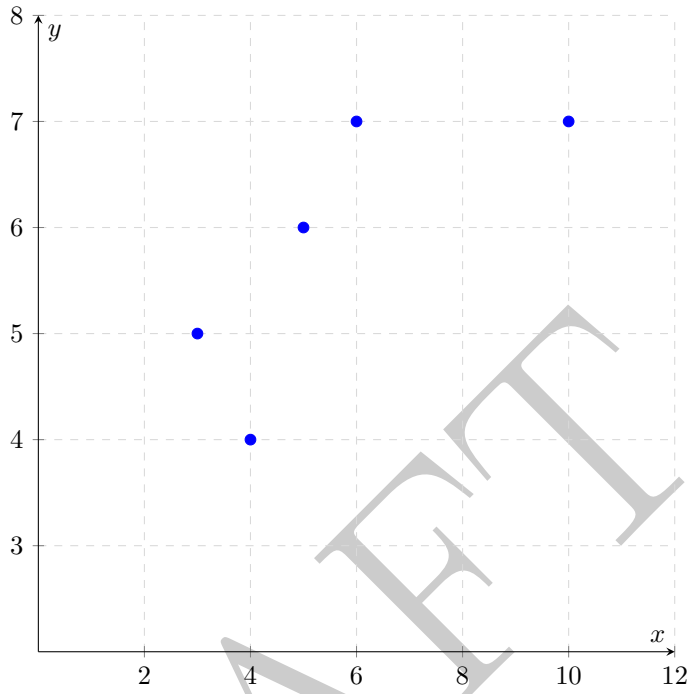


Figure 11.5: The scatter plot between x and y.

The scatter plot indicates that there is a positive correlation between X and Y, as the points generally trend upwards. This indicates that higher values of X tend to be associated with higher values of Y. However, to draw definitive conclusions about the strength and nature of this relationship, further statistical analysis, such as calculating the Pearson correlation coefficient, would be necessary.

(b)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
4	4	16	16	16
5	6	25	36	30
3	5	9	25	15
6	7	36	49	42
10	7	100	49	70
$\sum x_i = 28$	$\sum y_i = 29$	$\sum x_i^2 = 186$	$\sum y_i^2 = 175$	$\sum x_i y_i = 173$
$\bar{x} = 5.6$	$\bar{y} = 5.8$			

DRAFT

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} \\
&= \frac{173 - 5 \times 5.6 \times 5.8}{\sqrt{186 - 5 \times (5.6)^2}\sqrt{175 - 5 \times (5.8)^2}} \\
&= 0.7522
\end{aligned}
$$

The value $r = 0.7522$ indicates a strong positive linear relationship between the two variables X and Y. In general, values of r between 0.7 and 1.0 (or -0.7 and -1.0 for negative relationships) are considered to represent a strong relationship.

11.6.3 Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient involves determining whether the observed correlation between two variables is statistically significant, meaning it is unlikely to have occurred by chance. Here's a step-by-step guide on how to perform this test:

1. **Hypothesis**

   $H_0 : \rho = 0$ (There is no linear relations between two variables.)
   $H_1 : \rho \neq 0$ (There is a linear relations between two variables.)

2. **Level of significance** α

3. **Test statistic**

   $$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t \text{ distribution with } n-2 \text{ degrees of freedom}$$

4. **Decision Rule**
   Reject H_0 if

   $$T > t_{\alpha/2,n-2} \quad \text{or} \quad T < -t_{\alpha/2,n-2}$$

   equivalently,

   $$\text{Reject } H_0 \text{ if and only if } |T| > t_{\alpha/2,n-2}$$

   at level of significance $100 \times (1-\alpha)\%$.

Correlation Test: In general, if the null hypothesis is

$$H_0 : \rho = 0$$

and if the null hypothesis is true, the test statistic T follows the Student's-t

DRAFT

distribution with $(n-2)$ degrees of freedom, i.e., $T \sim t(n-2)$.

Table 11.5: Decision rule for the test of hypothesis $H_0 : \rho = 0$

Alternative hypothesis	Reject H_0 if
$H_1 : \rho < 0$	$T < -t_{\alpha,n-2}$
$H_1 : \rho > 0$	$T > t_{\alpha,n-2}$
$H_1 : \rho \neq 0$	$T > t_{\alpha/2,n-2}$ or $T < -t_{\alpha/2,n-2}$

For example given in Problem 11.3, we consider the realized value of T is t_{cal}. Therefore,

$$t_{\text{cal}} = \frac{0.7522 \times \sqrt{5-2}}{\sqrt{1-0.7522^2}} = 1.9772$$

The critical value is $t_{0.025,3} = 3.182$. Since, $|t_{\text{cal}}| = |1.9772| < 3.182$, we do not reject the null hypothesis that $\rho = 0$ and conclude that there is not enough evidence of a linear relationship between x and y in the population.

11.6.4 Python Code: Correlation Matrix

```
# example 1: using scipy library
from scipy.stats import pearsonr

# Data
heights = [150, 160, 170, 180, 190, 182]  # Heights in cm
weights = [55, 60, 65, 70, 75, 81]        # Weights in kg

# Calculate Pearson correlation coefficient
correlation, p_value = pearsonr(heights, weights)

print(f"Pearson correlation coefficient: {correlation}")
print(f"P-value: {p_value}")

# Example 2: for pandas dataset
  import pandas as pd

  # Example data (you should replace this with your actual
   data)
  data = pd.DataFrame({
    'x1': [1, 2, 3, 4, 5],
    'x2': [2, 3, 4, 5, 6],
    'x3': [3, 4, 5, 6, 7]
  })

  # Compute correlation matrix
  correlation_matrix = data.corr()
```

```
26
27      print(correlation_matrix)
28      # Compute correlation matrix with 2 decimal places
29      correlation_matrix = data.corr().round(2)
30
31      # View the correlation matrix
32      print(correlation_matrix)
33
34
```

11.7 Rank Correlation

Rank correlation is a statistical measure used to evaluate the strength and direction of the association between two ranked variables. Unlike Pearson's correlation coefficient, which assesses linear relationships between two continuous variables, rank correlation methods focus on the order or ranking of the data rather than the actual numerical values.

11.7.1 Key Types of Rank Correlation

Spearman's Rank Correlation Coefficient (Spearman's rho, ρ)

Spearman's rank correlation coefficient is a non-parametric measure of rank correlation. It is measured the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable. It assesses how well the relationship between two variables can be described using a monotonic function.

If r_i, s_i are the ranks of the i-member according to the x and the y-quality respectively, then the rank correlation coefficient is

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of data points of the two variables and d_i is the difference in the ranks of the ith element of each random variable considered, i.e. $d_i = r_i - s_i$, is the difference between ranks.

The Spearman correlation coefficient, ρ, can take values from $+1$ to -1.

- $\rho = 1$: Perfect positive correlation.
- $\rho = -1$: Perfect negative correlation.
- $\rho = 0$: No correlation.

DRAFT

Problem 11.4. *Based on the following data, find the rank correlation between marks of English and Mathematics courses.*

English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

Solution

The procedure for ranking these scores is as follows:

English (mark)	Maths (mark)	Rank (Engilsh)	Rank (Maths)	d_i	d_i^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	-3	9
62	65	6	5	1	1
64	56	5	9	-4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	-1	1
61	63	7	6	1	1

The realized value of the Spearman Rank Correlation is

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 54}{10(10^2-1)} = 0.6727$$

This indicates a strong positive relationship between the ranks individuals obtained in the Maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Equal Ranks or Tie in Ranks

- For **tie** observations, the rank correlation can be computed by adjusting $m(m^2-1)/12$ to the value of $\sum d_i^2$, where m stands for the number of items whose ranks are equal.
- If there are more than one such group of items with common rank, this value is added as many times as the number of such groups.

DRAFT

- Then the formula for the rank correlation is

$$\rho = 1 - \frac{6\{\sum d_i^2 + \frac{m_1}{12}(m_1^2 - 1) + \frac{m_2}{12}(m_2^2 - 1) + \cdots\}}{n(n^2 - 1)}$$

Problem 11.5. *Based on the following data, Find the rank correlation between marks of English and Mathematics courses.*

English	56	75	45	71	61	64	58	80	76	61
Maths	70	70	40	60	65	56	59	70	67	80

Solution

The procedure for ranking these scores is as follows:

English (mark)	Maths (mark)	Rank (Engilsh)	Rank (Maths)	d_i	d_i^2
56	70	9	3	6	36
75	70	3	3	0	0
45	40	10	10	0	0
71	60	4	7	-3	9
61	65	6.5	6	0.5	0.25
64	56	5	9	-4	16
58	59	8	8	0	0
80	70	1	3	-2	4
76	67	2	5	-3	9
61	80	6.5	1	5.5	30.25

- The mark 61 is repeated 2 times in series X (in English) and hence $m_1 = 2$.
- In series Y (in Math), the marks 70 occurs 3 times and hence $m_2 = 3$. So the rank correlation is

$$\begin{aligned}\rho &= 1 - \frac{6\{104.5 + \frac{2}{12}(2^2 - 1) + \frac{3}{12}(3^2 - 1)\}}{10(10^2 - 1)} \\ &= 0.3569\end{aligned}$$

DRAFT

11.7.2 Applications of Rank Correlation

- **Non-Linear Relationships**: Rank correlation is useful when the relationship between variables is not linear but still monotonic (e.g., one variable consistently increases as the other does, but not necessarily in a straight line).
- **Ordinal Data**: It is appropriate for ordinal data, where the values represent rankings or ordered categories (e.g., customer satisfaction ratings).
- **Handling Outliers**: Since rank correlation relies on the order of values rather than their magnitude, it is less sensitive to outliers than Pearson's correlation.

11.7.3 Python Code: Rank Correlation

In Python, you can find rank correlation using the scipy.stats module, which provides a function called spearmanr() to calculate the Spearman rank correlation coefficient. Here's how you can do it:

```
import scipy.stats as stats

# Given data
english_marks = [56, 75, 45, 71, 61, 64, 58, 80, 76, 61]
maths_marks = [70, 70, 40, 60, 65, 56, 59, 70, 67, 80]

# Calculate the Spearman rank correlation coefficient
spearman_corr, p_value = stats.spearmanr(english_marks, maths_marks)

# Output the results
print(f"Spearman Rank Correlation Coefficient: {spearman_corr:.4f}")
print(f"P-value: {p_value:.4f}")

```

If you have two columns of data in a pandas DataFrame, you can calculate the rank correlation directly from the DataFrame. Here's an example:

```
    # From Data file
    import pandas as pd
    # Example DataFrame (you should replace this with your actual data)
    data = pd.DataFrame({
      'english_marks': [56, 75, 45, 71, 61, 64, 58, 80, 76, 61],
      'maths_marks':  [70, 70, 40, 60, 65, 56, 59, 70, 67, 80]
    })
```

```
8
9       # Calculate Spearman rank correlation coefficient
10      rho, p_value = data.corr(method='spearman').iloc[0, 1]
11
12      print("Spearman rank correlation coefficient:", rho)
13      print("p-value:", p_value)
14
15
```

11.7.4 Kendall Tau Correlation Coefficient

Kendall's tau is another non-parametric measure of rank correlation that evaluates the ordinal association between two variables.

Calculation of Kendall Tau

To calculate the Kendall Tau correlation coefficient:

1. Compare each pair of observations in terms of their ranks.
2. Determine whether they have the same order (concordant) or opposite order (discordant) for both variables.
3. Count the total number of concordant pairs (n_c) and discordant pairs (n_d).
4. Calculate the Kendall Tau coefficient using the formula:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

5. It ranges from -1 to 1.

Interpretation:

- $\tau = 1$: Perfect agreement between the rankings.
- $\tau = -1$: Perfect disagreement between the rankings.
- $\tau = 0$: No association.

- **Concordant Pairs:**

  In the context of correlation coefficients such as Kendall Tau, concordant pairs refer to pairs of observations where the ranks for both variables follow the same order. In other words, if (X_i, Y_i) and (X_j, Y_j) are two pairs of observations, they are considered concordant if both $X_i < X_j$ and $Y_i < Y_j$ or if both $X_i > X_j$ and $Y_i > Y_j$.

DRAFT

- **Discordant Pairs:** Discordant pairs, on the other hand, refer to pairs of observations where the ranks for the variables have opposite orders. In other words, if (X_i, Y_i) and (X_j, Y_j) are two pairs of observations, they are considered discordant if $X_i < X_j$ and $Y_i > Y_j$ or if $X_i > X_j$ and $Y_i < Y_j$.

In the context of calculating correlation coefficients like Kendall Tau, understanding concordant and discordant pairs is crucial as they form the basis for determining the strength and direction of association between two variables based on their ranks.

Problem 11.6. *Suppose we have the following data on two variables, X and Y, with their corresponding ranks:*

Observation	*Rank*
X	Y
10	15
15	10
20	20
25	25
30	40

Calculate the Kendall Tau correlation coefficient.

Solution

To calculate the Kendall Tau correlation coefficient, we first need to identify the number of concordant and discordant pairs from the given data.

The data provided is:

Observation	Rank
X	Y
10	15
15	10
20	20
25	25
30	40

Let's denote the observations as (X_i, Y_i), where i ranges from 1 to 5. We compare each pair of observations (X_i, Y_i) with (X_j, Y_j) where $i < j$. Let's list all possible pairs and classify them:

DRAFT

1. Comparing (10, 15) with subsequent pairs:
    - (10, 15) vs (15, 10): $X_1 < X_2$ (10 < 15) but $Y_1 > Y_2$ (15 > 10). This is a **discordant** pair.
    - (10, 15) vs (20, 20): $X_1 < X_3$ (10 < 20) and $Y_1 < Y_3$ (15 < 20). This is a **concordant** pair.
    - (10, 15) vs (25, 25): $X_1 < X_4$ (10 < 25) and $Y_1 < Y_4$ (15 < 25). This is a **concordant** pair.
    - (10, 15) vs (30, 40): $X_1 < X_5$ (10 < 30) and $Y_1 < Y_5$ (15 < 40). This is a **concordant** pair.
2. **Comparing (15, 10) with subsequent pairs:**
    - (15, 10) vs (20, 20): $X_2 < X_3$ (15 < 20) and $Y_2 < Y_3$ (10 < 20). This is a **concordant** pair.
    - (15, 10) vs (25, 25): $X_2 < X_4$ (15 < 25) and $Y_2 < Y_4$ (10 < 25). This is a **concordant** pair.
    - (15, 10) vs (30, 40): $X_2 < X_5$ (15 < 30) and $Y_2 < Y_5$ (10 < 40). This is a **concordant** pair.
3. **Comparing (20, 20) with subsequent pairs:**
    - (20, 20) vs (25, 25): $X_3 < X_4$ (20 < 25) and $Y_3 < Y_4$ (20 < 25). This is a **concordant** pair.
    - (20, 20) vs (30, 40): $X_3 < X_5$ (20 < 30) and $Y_3 < Y_5$ (20 < 40). This is a **concordant** pair.
4. **Comparing (25, 25) with subsequent pairs:**
    - (25, 25) vs (30, 40): $X_4 < X_5$ (25 < 30) and $Y_4 < Y_5$ (25 < 40). This is a **concordant** pair.

So, out of the 10 pairs of observations, there are 9 concordant pairs and 1 discordant pair.

- Number of concordant pairs (n_c): 9
- Number of discordant pairs (n_d): 1
- Total number of pairs (n): $\frac{5(5-1)}{2} = 10$
- Kendall Tau coefficient (τ):

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} = \frac{9-1}{\frac{1}{2}(5)(5-1)} = \frac{8}{10} \approx 0.8$$

The Kendall Tau correlation coefficient for the given data is $\tau = 0.8$.

DRAFT

11.7.5 Advantages and Disadvantages

Rank correlation methods, such as Spearman's rank correlation coefficient and Kendall's tau are especially useful when dealing with ordinal data or when the assumptions of parametric methods (like Pearson's correlation) are not met. Here are some key advantages and disadvantages of rank correlation:

- **Advantages**:
  - **Robust to Outliers**: Rank correlation is not affected by extreme values.
  - **Non-parametric**: No assumptions about the distribution of the data.
  - **Suitable for Ordinal Data**: Can be used when data are ranked or ordinal.
- **Disadvantages**:
  - **Less Sensitive**: Might not capture the strength of the relationship as well as Pearson's correlation in the presence of linear relationships.
  - **Reduced Power**: May have less statistical power compared to parametric tests when assumptions for those tests are met.

11.7.6 Python Code: Kendall Tau

```
import scipy.stats as stats

# Given data
X = [10, 15, 20, 25, 30]
Y = [15, 10, 20, 25, 40]

# Calculate the Kendall Tau correlation coefficient
tau, p_value = stats.kendalltau(X, Y)

# Output the results
print(f"Kendall Tau correlation coefficient: {tau:.4f}")
print(f"P-value: {p_value:.4f}")
```

11.7.7 Exercises

1. Define a scatter diagram. What is its primary purpose in data analysis, and how can it help in understanding the relationship between two variables?

DRAFT

2. Describe how you would interpret a scatter diagram that shows a perfect positive linear relationship between two variables. What characteristics would you expect to see in the plot?

3. Explain how a scatter diagram can be used to detect non-linear relationships between variables. Provide examples of different types of non-linear relationships that might be observed.

4. What are some limitations of using a scatter diagram for analyzing relationships between variables? How can these limitations affect the interpretation of the data?

5. Define covariance. What does it measure, and how is it different from correlation?

6. What is meant by correlation analysis? How do you interpret the value of correlation coefficient?

7. Define Pearson's correlation coefficient. What does it measure, and how is it calculated?

8. Describe the range of values that Pearson's correlation coefficient can take and what each value signifies about the relationship between two variables.

9. Explain the concept of linearity in relation to Pearson's correlation coefficient. Why is Pearson's correlation coefficient not appropriate for non-linear relationships?

10. Discuss the assumptions underlying Pearson's correlation coefficient. What assumptions need to be met for Pearson's correlation coefficient to provide a valid measure of the strength and direction of the relationship?

11. How does Pearson's correlation coefficient handle outliers in the data? What impact can outliers have on the correlation coefficient, and how might this affect data analysis?

12. What is the purpose of using rank correlation methods, such as Spearman's and Kendall's Tau, instead of Pearson's correlation coefficient?

13. Explain the differences between Spearman rank correlation and Kendall Tau correlation. In what situations might one method be preferred over the other?

14. Discuss the advantages and limitations of using rank-based methods for measuring the association between two variables.

15. Define Spearman's rank correlation coefficient. How is it computed, and what does it measure?

DRAFT

16. Describe how tied ranks are handled in the computation of Spearman's rank correlation coefficient. What impact do ties have on the correlation value?

17. Define Kendall's Tau correlation coefficient. How does it differ from Spearman's rank correlation in terms of interpretation and calculation?

18. Explain the concepts of concordant and discordant pairs in the context of Kendall's Tau. How are they used to calculate the correlation coefficient?

19. Kendall's Tau is often considered more robust than Spearman's rank correlation in the presence of tied ranks. Discuss why this is the case and how Kendall's Tau adjusts for ties.

20. Given the following pairs of data representing the number of hours studied (X) and the scores obtained (Y) by 10 students in an exam:

Student	Hours Studied (X)	Exam Score (Y)
1	2	50
2	3	55
3	5	60
4	6	65
5	8	70
6	9	75
7	10	80
8	12	85
9	14	90
10	15	95

(i). Plot the scatter diagram for this data. Describe the relationship between the hours studied and the exam scores based on the scatter plot.

(ii). Calculate the covariance between the number of hours studied (X) and the exam scores (Y).

(iii). Interpret the sign and magnitude of the covariance. What does it tell you about the relationship between the two variables?

(iv). Calculate the Pearson correlation coefficient between the hours studied and exam scores.

DRAFT

(v). Interpret the value of the correlation coefficient. What does it indicate about the strength and direction of the relationship between the two variables?

21. The owner of Maumee Ford-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Car	Age (years)	Selling Price ($000))
1	9	8.1
2	7	6.0
3	11	3.6
4	12	4.0
5	8	5.0
6	7	10.0
7	8	7.6
8	11	8.0
9	10	8.0
10	12	6.0
11	6	8.6
12	6	8.0

(i). Draw a scatter diagram. Interpret the plot.

(ii). Determine the correlation coefficient.

(iii). Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

(iv). Test whether this linear relationship is significant or not.

22. Consider the following data on the ranks of two variables, X and Y:

Observation	Rank of X	Rank of Y
1	2	3
2	1	2
3	4	1
4	3	4

(a) Calculate the Spearman rank correlation coefficient for the above data.

DRAFT

(b) Interpret the result in the context of the relationship between X and Y.

23. Suppose we have the following ranks for two variables A and B:

Observation	Rank of A	Rank of B
1	1	2
2	2	1
3	3	3
4	4	5
5	5	4

(a) Calculate the Kendall Tau correlation coefficient for the data provided.

(b) Discuss the strength and direction of the relationship between A and B based on your result.

24. You are provided with the following data on two variables, P and Q. Compute both the Spearman rank correlation and Kendall Tau correlation coefficients:

Observation	P	Q
1	85	92
2	78	85
3	92	88
4	70	76
5	88	90

(a) Compare the results from both rank correlation methods. Explain any similarities or differences observed.

11.8 Regression Analysis

Imagine you are a data scientist working for a real estate company. Your task is to predict the price of a house based on various features such as the number of bedrooms, square footage, and location. You have a dataset with historical sales data, and you want to use this data to build a model that can predict the price of a house given its features. This is where regression analysis comes into play.

The drawback of correlation analysis is that it only measures the strength and direction of the (linear) relationship between two variables. It does not

DRAFT

provide information about **causality** or the **nature of the relationship** beyond linearity. Additionally, correlation coefficients can be affected by outliers and may not capture complex relationships that exist between variables.

The motivation for regression analysis stems from the need to understand and model the relationship between variables more comprehensively. Regression analysis allows us to not only measure the direction and significant effect of the relationship but also to make **predictions** and **infer causality**, provided certain assumptions are met. By fitting a regression model, we can examine how changes in one variable are associated with changes in another variable while controlling for potential confounding factors.

- **Dependent Variable:** The variable that is the outcome or response being predicted or explained in an analysis.
- **Independent Variable:** The variable that is the predictor or factor used to predict or explain changes in the dependent variable.

Regression analysis enables deeper insights into the underlying mechanisms driving the relationship between variables. It is a set of statistical methods used to estimate relationships between a dependent variable (also known as the response or target variable) and one or more independent variables (also known as predictors, features, or explanatory variables), allowing for predictions and the determination of the strength and nature of these relationships. This relationship can be expressed as

$$y = g(x_1, x_2, \ldots, x_p) + e \tag{11.2}$$

where y represents the dependent variable, $x_1, x_2, \ldots, x_p$ are the independent variables, and e denotes the error term.

Regression analysis: Regression analysis is a statistical tool used to establish a mathematical relationship between a set of independent variables and a dependent variable. It helps us to (i) quantify the strength and direction of the relationships, (ii) assess the statistical significance of these relationships, (iii) understand how the response variable depends on the independent variables and how these factors work together, and (iv) enable prediction of future or unknown values.

In model (11.2), the function $g(x_1, x_2, \ldots, x_p)$ represents the systematic part of the model. It describes the relationship between the predictor variables $x_1, x_2, \ldots, x_p$ and the response variable y. The specific form of g depends on the type of regression model. For example, in **linear regression**, g is typically a linear combination of the predictors:

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

For instance, banks might use regression analysis to evaluate the risk associated with home-loan applicants. In this context, independent variables such as

the applicant's age, income, expenses, occupation, number of dependents, and total credit are used to predict the likelihood of loan repayment, thereby aiding in risk assessment and decision-making processes. The main goal is to understand how changes in the independent variables affect the dependent variable, and to use this understanding to make predictions.

Mean regression estimates the conditional mean of the response variable, answering questions about what happens "on average" for given predictor values. However, real-world relationships are often more nuanced. The effect of an independent variable may differ substantially across different parts of the outcome distribution. For instance, the impact of education on income might be stronger for high earners than for low earners, or the effect of a medical treatment might vary between patients with mild versus severe symptoms.

This is where **quantile regression** becomes invaluable. Instead of summarizing the relationship with a single line through the center of the data, quantile regression allows us to estimate relationships at any percentile of the conditional distribution—the 25th, 50th (median), 75th, or 90th percentile, for example. This approach is particularly robust to outliers and provides a more complete picture of how variables relate across the entire spectrum of outcomes, making it especially useful when studying inequality, risk assessment, or situations where tail behaviors matter more than averages. In this chapter, we focus exclusively on mean regression.

11.8.1 Types of regression analysis

In a broad sense, there are two types of regression models: (i) parametric regression models and (ii) nonparametric regression models. Parametric regression involves assuming a specific functional form for the relationship between the dependent variable and the independent variables. The model is defined by a finite number of parameters. The most common example is linear regression, where the relationship between the variables is assumed to be linear. Non-parametric regression does not assume a specific functional form for the relationship between the dependent and independent variables. Instead, the nonparametric regression model is more flexible and can adapt to the data's structure. We will not discuss nonparametric regression models; instead, we will focus solely on parametric regression models.

There are various types of regression models depending on the nature of the response variable. Some of them are mentioned below:

- for continuous response variable
  - Simple Linear Regression
  - Multiple Regression

DRAFT

- Polynomial Regression
- Multivariate Regression, etc.

- for categorical response variable
  - Logistic Regression
    - binomial (also called ordinary) logistic regression
    - multinomial logistic regression
    - ordinal logistic regression
    - alternating logistic regressions, etc.

- for discrete response variable
  - Poisson Regression
  - Negative Binomial Regression, etc.

- for survival-time (time-to-event) outcomes
  - Cox Regression (or proportional hazards regression)
  - Accelerated Failure Time Model (AFT model), etc.

11.8.2 Simple Regression Model

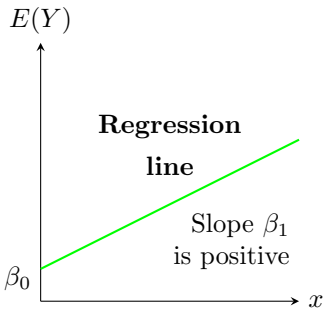
To establish the relationship between the average value of the dependent variable Y given each value of the independent variable X, we consider the following **population regression function**:

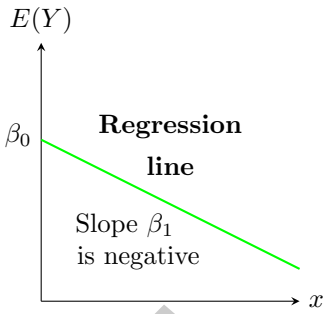
$$E(Y \mid X = x) = \beta_0 + \beta_1 x \tag{11.3}$$

where

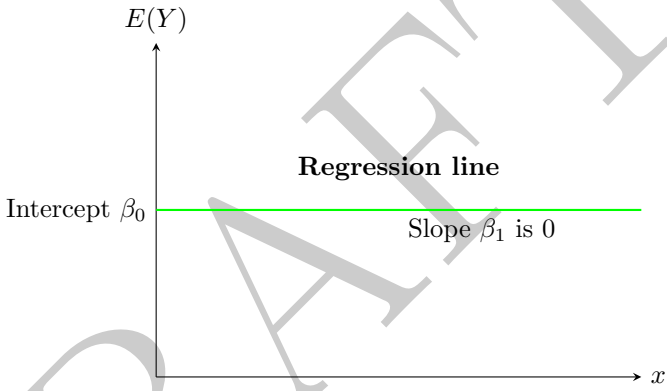
- $E(Y \mid X = x)$ represents the expected (or average) value of Y when X is equal to x.
- β_0 and β_1 are the parameters of the regression function, with β_0 being the **intercept** and β_1 the **slope** of the regression line.

DRAFT





Panel A: Positive Linear Relationship Panel B: Negative Linear Relationship



Panel C: No Relationship

Suppose x_i is a given value of X for the i-th observation. Then the population regression function (PRF) can be written as

$$E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i \tag{11.4}$$

where $E(Y \mid X = x_i)$ represents the expected value of Y given that X takes the value x_i. For a given value $x = x_i$, the observed value y_i for Y can be expressed as:

$$\begin{aligned} y_i &= E(Y \mid X = x_i) + e_i \\ &= \beta_0 + \beta_1 x_i + e_i \end{aligned} \tag{11.5}$$

where e_i is the error term for the i-th observation, representing the deviation of the observed value y_i from the expected value $E(Y \mid X = x_i)$. The model given by Equation (11.5) is known as the **simple linear regression model**. This model is used to estimate the population regression function described in Equation (11.4).

DRAFT

The **classical simple linear regression (CSLR) model**, which follows the assumptions mentioned in Section 11.8.3, is defined for the i-th observation as

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, n \tag{11.6}$$

where:

- y_i is the dependent variable (response) for observation i.
- x_i is the independent variable (predictor) for observation i.
- β_0 is the intercept of the regression line.
- β_1 is the slope of the regression line.
- e_i is the error term (residual) for observation i, representing the difference between the observed and predicted values.

Note that the model (11.6) is a special case of (11.2) when g is a linear function and $p = 1$.

A **nonintercept model** in regression analysis is a model where the regression line or hyperplane is forced to pass through the origin (the point where all variables are zero). This means the intercept term is explicitly set to zero and is not estimated from the data. The model equation simplifies to:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

where y is the dependent variable, x_i are the independent variables, β_i are the coefficients, and e is the error term. There is no constant term (β_0).

For instance, in a manufacturing process, if the number of products made depends directly on the amount of raw material used, and no products can be produced without any raw material, a non-intercept model is appropriate. The model would be

$$\text{Products} = \beta_1 \times \text{Raw Material} + e$$

where β_1 estimates the production rate per unit of raw material. This model is suitable when there's a clear theoretical reason to exclude the intercept, but caution is needed if a baseline effect exists.

11.8.3 Assumptions of the CSLR Model (11.6)

The assumptions of the CSLR model given in Equation (11.6) are as follows:

(i). **Linearity:** The relationship between X and Y must be linear.

(ii). **Independence of Errors:** The errors should be independent of each other. More precisely, there should be no correlation between the errors and the independent variable X.

DRAFT

(iii). **Zero Mean:** For any fixed value of x_i, the mean of the errors (residuals) is zero. That is, $E(e_i \mid x_i) = 0$.

(iv). **Homoscedasticity:** The variance of the errors (residuals) is constant across all levels of the independent variable X. That is, for any fixed value of x_i, $\text{Var}(e_i \mid x_i) = \sigma^2$.

(v). **Normality:** The errors are normally distributed for any fixed value of x_i.

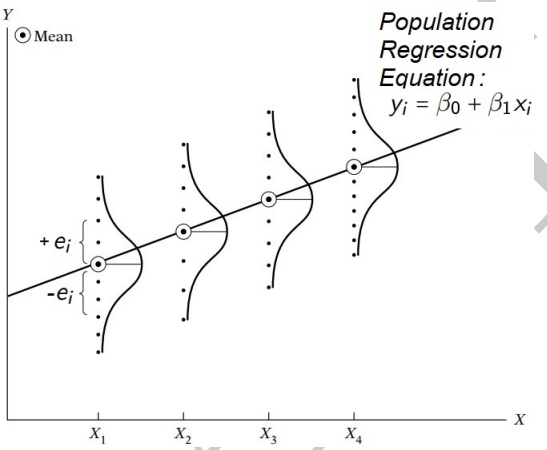


Figure 11.6: Conditional distribution of the disturbances e_i

Under these assumptions, we can write $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, and hence, $y_i \mid x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Therefore, this model is also known as a **normal linear regression** model. The graphical representation of the simple linear regression model given in Equation (11.6) under these assumptions is presented in Figure 11.6.

Remark 11.8.1. *Normality of the errors (or residuals) is not strictly required. However, the normality assumption in Equation* (11.6) *is necessary to perform hypothesis tests concerning regression parameters, as discussed in the next.*

The **estimated model** (or **estimated line**) for Equation (11.6) can be written as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where

- $\widehat{y}_i$ is the estimator of $E(Y \mid X = x_i)$ based on the sample data.
- $\widehat{\beta}_0$ is the estimator of β_0.

- $\widehat{\beta}_1$ is the estimator of β_1.

Therefore, in terms of the sample regression function (SRF), the observed value y_i can be expressed as

$$y_i = \widehat{y}_i + \widehat{e}_i$$

where $\widehat{e}_i$ is the residual, representing the deviation of the observed value y_i from the estimated value $\widehat{y}_i$.

11.8.4 Ordinary Least Squares (OLS) Estimation

Ordinary Least Squares (OLS) is a method for estimating the parameters in a linear regression model. The basic idea behind OLS is to find the best line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences (error) between the observed values y and the predicted values of y from the linear model.

Under the assumptions given in Section 11.8.3 for the simple linear regression model provided in Equation (11.6), the OLS estimators for the parameters β_0 and β_1 are found by minimizing the sum of squared errors. Thus, the objective function for OLS estimation is:

$$\text{Minimize} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

where $\sum_{i=1}^{n} e_i^2$ is the sum of squared errors.

OLS Estimators

The OLS estimators for β_0 and β_1 can be derived by taking the partial derivatives of $Q(\beta_0, \beta_1)$ with respect to β_0 and β_1, setting them to zero, and solving for the coefficients. The OLS estimators are:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$
$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where:

- $\widehat{\beta}_1$ is the estimated slope of the regression line.
- $\widehat{\beta}_0$ is the estimated intercept of the regression line.
- $\bar{x}$ and $\bar{y}$ are the sample means of the independent variable X and the dependent variable Y, respectively.

DRAFT

Procedure for Calculating the OLS Estimators in a Simple Linear Regression Model

For the simple linear regression model given in Equation (11.6), the OLS estimators can be found using the following steps:

1. **Objective Function**:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

2. **Find the Estimators**: The estimators for β_0 and β_1 are obtained by solving the following normalized equations:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{11.7}$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} e_i^2 = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \tag{11.8}$$

From Equation (1), we get:

$$\begin{aligned} &\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\ &\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x} \end{aligned}$$

From Equation (2), we get:

$$\begin{aligned} &\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\ &\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \end{aligned} \tag{11.9}$$

DRAFT

Substitute $\beta_0 = \bar{y} - \beta_1 \bar{x}$ into Equation (3):

$$\begin{aligned}
&\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \beta_1 \bar{x} \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} - \beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = 0 \\
&\Rightarrow \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}
\end{aligned}$$

Hence, the **Ordinary Least Squares (OLS) estimators** are:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x}
\end{aligned}$$

The estimator $\widehat{\beta}_1$ can also be expressed as:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r \left(\frac{s_y}{s_x} \right)$$

where r is the correlation coefficient between X and Y, and s_x and s_y are the sample standard deviations of X and Y, respectively.

11.8.5 Interpretation of Regression Coefficients

The estimated regression model (or equation) is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are respectively, the estimator of β_0 and β_1

- **Intercept ($\widehat{\beta}_0$)**: The intercept represents the expected value of the dependent variable Y when the realized value of independent variable X is zero. In other words, it is the point where the regression line crosses the y-axis.
- **Slope of the line ($\widehat{\beta}_1$)**:
  - it shows the **amount of change** in $\widehat{y}$ for a **change of one unit** in X
  - a positive value for $\widehat{\beta}_1$ indicates a direct relationship between the two variables and a negative value indicates an inverse relationship

DRAFT

- the sign of $\widehat{\beta}_1$ and the sign of r, the correlation coefficient, are **always the same**

Important properties related to correlation and regression coefficients are stated in Theorem 11.1.

Theorem 11.1. *Let $\widehat{\beta}_{xy}$ be the slope of the regression of y on x and $\widehat{\beta}_{yx}$ be the slope of the regression of x on y. The estimated regression lines are:*

- $y = \widehat{\beta}_{xy}x + b_{xy}$ *(regression of y on x)*
- $x = \widehat{\beta}_{yx}y + b_{yx}$ *(regression of x on y)*

Then the geometric mean of $\widehat{\beta}_{xy}$ and $\widehat{\beta}_{yx}$ is equal to the absolute value of the Pearson correlation coefficient r.

Proof. The slope $\widehat{\beta}_{xy}$ of the regression of y on x is given by:

$$\widehat{\beta}_{xy} = r \cdot \frac{s_y}{s_x}$$

where r is the Pearson correlation coefficient, s_y is the standard deviation of y, and s_x is the standard deviation of x.

The slope $\widehat{\beta}_{yx}$ of the regression of x on y is given by:

$$\widehat{\beta}_{yx} = r \cdot \frac{s_x}{s_y}$$

To find the geometric mean of $\widehat{\beta}_{xy}$ and $\widehat{\beta}_{yx}$, we compute:

$$\text{Geometric Mean} = \sqrt{\widehat{\beta}_{xy} \cdot \widehat{\beta}_{yx}}$$

Substitute the expressions for $\widehat{\beta}_{xy}$ and $\widehat{\beta}_{yx}$:

$$\text{Geometric Mean} = \sqrt{\left(r \cdot \frac{s_y}{s_x}\right) \cdot \left(r \cdot \frac{s_x}{s_y}\right)} = \sqrt{r^2} = |r|$$

Thus, the geometric mean of the regression coefficients $\widehat{\beta}_{xy}$ and $\widehat{\beta}_{yx}$ is the absolute value of the Pearson correlation coefficient r. □

11.8.6 The Estimated Error Variance or Standard Error

The error variance or standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression. The formulas that are used to compute the error variance or the standard error:

$$\widehat{\sigma}^2 = \frac{\sum_i^n (y_i - \widehat{y}_i)^2}{n-2} = \frac{\sum_i^n y_i^2 - \widehat{\beta}_0 \sum_i^n y_i - \widehat{\beta}_1 \sum_i^n x_i y_i}{n-2}$$

DRAFT

where

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Hence, the **standard error of estimate** is

$$\widehat{\sigma} = \sqrt{\frac{\sum\limits_i (y_i - \widehat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum\limits_i^n y_i^2 - \widehat{\beta}_0 \sum\limits_i^n y_i - \widehat{\beta}_1 \sum\limits_i^n x_i y_i}{n-2}}$$

Problem 11.7 (Shorshe Ilish restaurant Sales Dataset)**.** *Suppose data were collected from a sample of 10 semesters from a restaurant located near to the university campus. For the ith semester in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars).*

Table 11.6: Student Population and Quarterly Sales Data for Different Semesters

Semesters i	**Student Population (1000s)** x_i	**Quarterly Sales ($1000s)** y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

(i). Show the relationship between the size of student population and the quarterly sales. Make a comment on the diagram.

(ii). Write down the regression model for this example, and mention the assumptions of the model.

(iii). Find the least square estimates and write the estimated regression model. Interpret the results.

(iv). Draw the regression line on the scatter diagram.

DRAFT

(v). Predict quarterly sales for a semesters to be located near a campus with 16,000 students.

(vi). Find the value of the standard error of the estimates.

Solution

(i). The scatter plot in Figure 11.7 shows the relationship between student population (in thousands) and quarterly sales (in thousands of dollars) based on data from ten semesters.

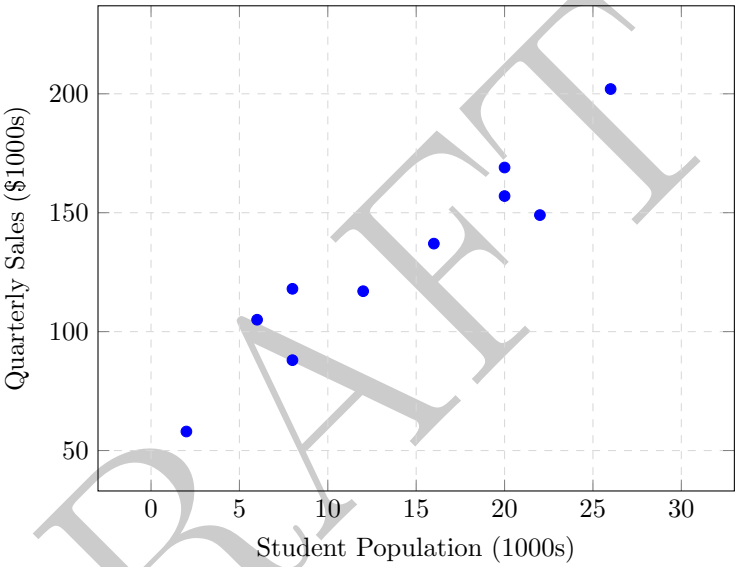


Figure 11.7: Scatter Plot of Student Population vs. Quarterly Sales

Observations:

- There appears to be a positive correlation between student population and quarterly sales. As the student population increases, the quarterly sales also tend to increase.
- The data points are somewhat clustered along a line, suggesting a linear relationship.
- Some points, such as the one where the student population is 8, show variability in sales, indicating that factors other than student population might also influence sales.

DRAFT

(ii). The regression model for this example is a simple linear regression model, which can be expressed as:

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{11.10}$$

- y_i represents the **quarterly sales** (in thousands of dollars) for the i-th semester.
- x_i represents the **student population** (in thousands) for the i-th semester.
- β_0 is the **intercept**, representing the expected quarterly sales when the student population is zero.
- β_1 is the **slope coefficient**, which represents the expected change in quarterly sales for every one-thousand-student increase in the student population.
- e_i is the **error term**, which accounts for the variability in sales that is not explained by the student population.

The **assumptions** of this model are:

(i). **Linearity**: The relationship between the student population and quarterly sales is linear.

(ii). **Independence of errors**: The error terms are independent of each other.

(iii). **Homoscedasticity**: The variance of the error term (σ^2) is constant for all levels of the student population.

(iv). **Normality of errors**: The error terms (e_i) are normally distributed.

(v). **Zero mean of errors**: The average value of the error term is zero.

(iii). The estimated regression model (or equation) of (11.10) is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are respectively, the estimator of β_0 and β_1. The **ordinary least square (OLS) estimators** of β_0 and β_1 are

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \quad ; \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

DRAFT

The Least Squares Estimation The **ordinary least square (OLS) estimators** of β_0 and β_1 are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{2840}{568} = 5$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 130 - 5(14) = 60$$

x_i	y_i	x_i^2	$x_i y_i$
2	58	4	116
6	105	36	630
8	88	64	704
8	118	64	944
12	117	144	1404
16	137	256	2192
20	157	400	3140
20	169	400	3380
22	149	484	3278
26	202	676	5252
Total= 140	1300	2528	21040

Thus, the estimated regression equation is

$$\widehat{y}_i = 60 + 5x_i$$

The slope of the estimated regression equation ($\widehat{\beta}_1 = 5$) is positive, implying that as student population increases, sales increase. In fact, we can conclude that an increase in the student population of 1000 is associated with an increase of \$5000 in expected sales; that is, quarterly sales are expected to increase by \$5 per student.

(iv). The regression line with the scatter plot is depicted in Figure 11.8.

DRAFT

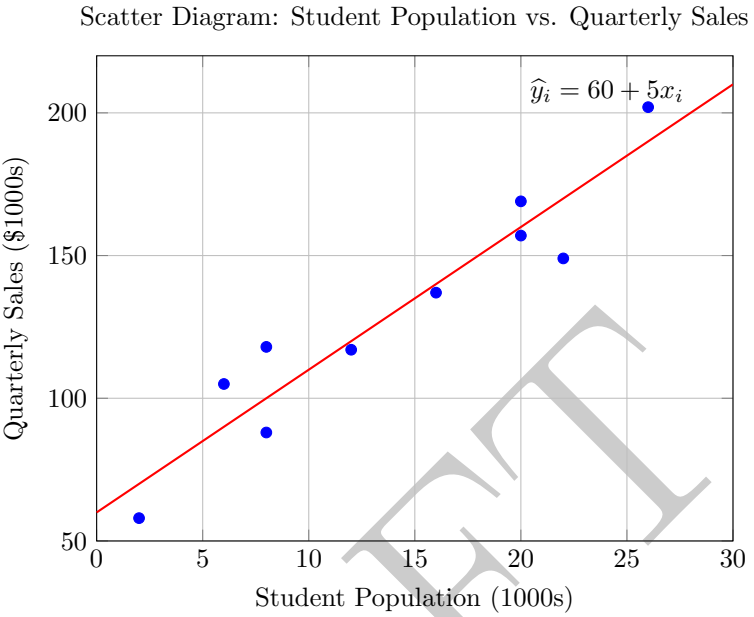


Figure 11.8: Scatter Plot of Student Population vs. Quarterly Sales with Regression Line

(v). To predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\widehat{y} = 60 + 5 \times 16 = 140$$

Hence, we would predict quarterly sales of $140,000 for this restaurant.

DRAFT

x_i	y_i	$\widehat{y}_i = 60 + 5x_i$	$(y_i - \widehat{y}_i)^2$
2	58	70	144
6	105	90	225
8	88	100	144
8	118	100	324
12	117	120	9
16	137	140	9
20	157	160	9
20	169	160	81
22	149	170	441
26	202	190	144
140	**1300**	**760**	**1530**

(vi). The Standard Error of the Estimates

$$\widehat{\sigma}^2 = \frac{\sum\limits_i (y_i - \widehat{y}_i)^2}{n-2} = \frac{1530}{10-2} = 191.25 \text{ and } \widehat{\sigma} = \sqrt{191.25} = 13.829$$

Hence the standard error of the estimates is $\widehat{\sigma} = 13.829$.

11.8.7 Coefficient of Determination

The **coefficient of determination** is the proportion of the total variation in the dependent variable Y that is explained by the independent variable(s) in a regression model. It is denoted by R^2 and defined by

$$R^2 = 1 - \frac{\sum\limits_i (y_i - \widehat{y}_i)^2}{\sum\limits_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

where,

- the **total sum of squares** (proportional to the variance of the data): $SST = \sum\limits_i (y_i - \bar{y})^2$
- the **sum of squares of residuals**, also called the **Residual or Error Sum of Squares (SSE)**: $SSE = \sum\limits_i (y_i - \widehat{y}_i)^2 = \sum\limits_i e_i^2$

For the **Shorshe Ilish restaurant Sales Dataset** given in Problem 11.7, we have

DRAFT

x_i	y_i	$(y_i - \bar{y})^2$	$\widehat{y}_i = 60 + 5x_i$	$(y_i - \widehat{y}_i)^2$
2	58	5184	70	144
6	105	625	90	225
8	88	1764	100	144
8	118	144	100	324
12	117	169	120	9
16	137	49	140	9
20	157	729	160	9
20	169	1521	160	81
22	149	361	170	441
26	202	5184	190	144
140	**1300**	**15730**	**1300**	**1530**

Hence,

$$R^2 = 1 - \frac{\sum\limits_i (y_i - \widehat{y}_i)^2}{\sum\limits_i (y_i - \bar{y})^2} = 1 - \frac{1530}{15730} = 0.9027$$

The $R^2 = 0.9027$ implies that 90.27% of the variability of the quarterly sales is explained and the remaining 9.73% of the variability is still unexplained by the regression model.

11.8.8 Relationship between R^2 and r

The relationship between R^2 and r is that the square of the coefficient of correlation (r) is equal to the coefficient of determination (R^2) for the **simple regression model**. Mathematically,

$$R^2 = (r)^2$$

Hence, for the previous Problem 11.3, it is given that

x	4	5	3	6	10
y	4	6	5	7	7

and we found

$$\begin{aligned} R^2 = (r)^2 &= (0.7522)^2 \\ &= 0.56580484 \approx 0.566 \end{aligned}$$

DRAFT

This means that 56.6% of the total variation in y is explained by the regression model if we fit a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$.

The estimated regression equation for simple linear regression model provides

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

The **sample correlation coefficient** is

$$\begin{aligned} r &= (\text{sign of } \widehat{\beta}_1)\sqrt{\text{coefficient of determination}} \\ &= (\text{sign of } \widehat{\beta}_1)\sqrt{R^2} \end{aligned} \tag{11.11}$$

where R^2 is the coefficient of determination for simple regression model $y_i = \beta_0 + \beta_1 x_i + e_i; \quad i = 1, 2, \ldots, n$

Remark 11.8.2. *Note that the relationship $R^2 = (r)^2$ only holds for the simple linear regression model.*

Problem 11.8. *Using the **Shorshe Ilish restaurant Sales Dataset** from Problem 11.7, calculate the following:*

(a) The Pearson correlation coefficient (r).

(b) The coefficient of determination (R^2).

(a). Pearson Correlation Coefficient

We will calculate the Pearson correlation coefficient for the **Shorshe Ilish restaurant Sales Dataset**. The data and necessary sums are presented in the following Table 11.7.

DRAFT

Table 11.7: Data and Sums for Correlation Calculation

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
2	58	4	3364	116
6	105	36	11025	630
8	88	64	7744	704
8	118	64	13924	944
12	117	144	13689	1404
16	137	256	18769	2192
20	157	400	24649	3140
20	169	400	28561	3380
22	149	484	22201	3278
26	202	676	40804	5252
$\sum x_i = 140$	$\sum y_i = 1300$	$\sum x_i^2 = 2528$	$\sum y_i^2 = 184730$	$\sum x_i y_i = 21040$
$\bar{x} = 14.0$	$\bar{y} = 130.0$			

The formula for the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Substituting the calculated values into the formula:

$$\begin{aligned} r &= \frac{21040 - 10 \times 14.0 \times 130.0}{\sqrt{(2528 - 10 \times (14.0)^2)}\sqrt{(184730 - 10 \times (130.0)^2)}} \\ &= \frac{21040 - 18200}{\sqrt{(2528 - 1960)}\sqrt{(184730 - 169000)}} \\ &= \frac{2840}{\sqrt{568}\sqrt{15730}} \\ &\approx \frac{2840}{2989.16} \\ &= 0.9503 \end{aligned}$$

The value of the Pearson correlation coefficient is $r = 0.9503$. This indicates a strong positive linear relationship between the student population and quarterly sales.

DRAFT

(b). The Coefficient of Determination (R^2)

For simple regression settings, we know, the coefficient of determination, R^2, is the square of the Pearson correlation coefficient (r). From our previous calculation in **(a)**, the Pearson correlation coefficient is $r \approx 0.9503$.

Therefore, the coefficient of determination is

$$R^2 = r^2 \approx (0.9503)^2 \approx 0.9027$$

This means that about 90.27% of the variation in quarterly sales can be explained by the linear relationship with the size of the student population.

11.8.9 Advantages and Disadvantages of R^2

Here are some advantages and disadvantages of using R^2:

- R^2 is a statistic that will give some information about the **goodness-of-fit** of a model.
- In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.
- An R^2 of 1 indicates that the regression predictions perfectly fit the data.
- R^2 increases as we increase the number of variables in the model (R^2 is **monotone increasing** with the number of variables included i.e., it will never decrease).

11.8.10 Adjusted R^2

An adjusted R^2 is a modification of R^2 that **adjusts** for the number of independent variables in a model (p) relative to the number of data points (n). The adjusted R^2 increases only if the new predictor improves the model more than would be expected by chance. It penalizes the model for each additional independent variable that does not contribute significantly to explaining the variance in the dependent variable. The adjusted R^2 (denoted by $\bar{R}^2$) is defined as

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p-1}\right).$$

where p is the total number of independent variables in the model (not including the constant term), and n is the sample size. It can also be written as:

$$\bar{R}^2 = 1 - \frac{SSE/\mathrm{df}_e}{SST/\mathrm{df}_t}$$

DRAFT

where df_t is the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance.

The explanation of this statistic is almost the same as R^2 but it **penalizes the statistic as extra variables** are included in the model. The term $\frac{n-1}{n-p-1}$ is called the **penalty** of using the more covariates in a model. When the number of covariates p, increases, $(1-R^2)$ will decrease, but $\frac{n-1}{n-p-1}$ will increase. Whether more covariates improve the explanatory power of a model depends on the **trade-of** between R^2 and the **penalty**.

11.8.11 Relationship between R^2 and $\bar{R}^2$ in Simple Linear Regression

For the simple linear regression model, we have $p=1$. Thus,

$$\bar{R}^2 = 1 - \left(1-R^2\right)\frac{n-1}{n-2}.$$

Since $\frac{n-1}{n-2} > 1$ for any finite $n > 2$, it follows that

$$\bar{R}^2 < R^2$$

unless $R^2 = 1$ (perfect fit). Therefore, in simple regression, the adjusted R^2 will be slightly less than the unadjusted R^2, with the difference becoming negligible as n becomes large.

For the previous Problem 11.3, $p=1$ and $R^2 = 0.566$ and hence,

$$\bar{R}^2 = 1 - (1-0.566)\left(\frac{5-1}{5-1-1}\right) = 0.4212$$

Python Code: Simple Regression Analysis

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Data: Student Population and Quarterly Sales
data = {
  'Restaurant': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
  'Student Population (1000s)': [2, 6, 8, 8, 12, 16, 20, 20, 22, 26],
  'Quarterly Sales ($1000s)': [58, 105, 88, 118, 117, 137, 157, 169, 149, 202]
```

DRAFT

```
}

# Convert the data to a DataFrame
df = pd.DataFrame(data)

# (i) Show the relationship between the size of the student population and the quarterly sales
plt.figure(figsize=(8, 6))
plt.scatter(df['Student Population (1000s)'], df['Quarterly Sales ($1000s)'], color='blue')
plt.title('Relationship between Student Population and Quarterly Sales')
plt.xlabel('Student Population (1000s)')
plt.ylabel('Quarterly Sales ($1000s)')
plt.grid(True)
plt.show()


# (iii) Write the estimated regression equation and find the least squares estimates
X = df['Student Population (1000s)']
y = df['Quarterly Sales ($1000s)']

# Add a constant to the independent variable matrix
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(y, X).fit()

# Print the regression results
print(model.summary())

# The estimated regression equation:
# y = model.params[0] + model.params[1] * x
# Interpret the results based on the model summary.

# (iv) Draw the regression line
plt.figure(figsize=(8, 6))
plt.scatter(df['Student Population (1000s)'], df['Quarterly Sales ($1000s)'], color='blue', label='Data Points')
plt.plot(df['Student Population (1000s)'], model.predict(X), color='red', label='Regression Line')
plt.title('Regression Line: Quarterly Sales vs. Student Population')
plt.xlabel('Student Population (1000s)')
plt.ylabel('Quarterly Sales ($1000s)')
plt.legend()
plt.grid(True)
plt.show()
```

```
53
54 # (v) Predict quarterly sales for a restaurant near a campus
       with 16,000 students
55 predicted_sales = model.predict([1, 16])[0]  # [1, 16] where
       1 is for the constant term
56 print(f'Predicted Quarterly Sales for 16,000 students: ${
       predicted_sales:.2f} (in $1000s)')
57
58 # (vi) Find the value of the standard error of the estimates
59 standard_error = np.sqrt(model.mse_resid)
60 print(f'Standard Error of the Estimates: {standard_error:.2f
       }')
61
62
```

Figure 11.9: Python output for the practice example.

```
================================================================================
Dep. Variable:     Quarterly Sales ($1000s)   R-squared:                       0.903
Model:                                 OLS   Adj. R-squared:                  0.891
Method:                      Least Squares   F-statistic:                     74.25
Date:                     Mon, 12 Aug 2024   Prob (F-statistic):           2.55e-05
Time:                             04:57:26   Log-Likelihood:                -39.342
No. Observations:                       10   AIC:                             82.68
Df Residuals:                            8   BIC:                             83.29
Df Model:                                1
Covariance Type:                 nonrobust
================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const                   60.0000      9.226      6.503      0.000      38.725      81.275
Student Population (1000s)     5.0000      0.580      8.617      0.000       3.662       6.338
==============================================================================
Omnibus:                        0.928   Durbin-Watson:                   3.224
Prob(Omnibus):                  0.629   Jarque-Bera (JB):                0.616
Skew:                          -0.060   Prob(JB):                        0.735
Kurtosis:                       1.790   Cond. No.                         33.6
==============================================================================
```

Problem 11.9 (Insulin and Blood Glucose Levels Dataset)**.** *Suppose data were collected from a sample of 12 patients with type 2 diabetes. For the ith patient, x_i is the amount of insulin administered (in units) and y_i is the corresponding fasting blood glucose level (in mg/dL).*

(i). Show the relationship between the amount of insulin administered and the fasting blood glucose level. Make a comment on the diagram.

(ii). Write down the regression model for this example, and mention the assumptions of the model.

(iii). Find the least square estimates and write the estimated regression model. Interpret the results.

Table 11.8: Insulin and Fasting Blood Glucose Data for Different Patients

Patient i	**Insulin Administered (units)** x_i	**Fasting Blood Glucose (mg/dL)** y_i
1	5	180
2	8	165
3	10	160
4	12	155
5	15	140
6	15	145
7	18	130
8	20	125
9	22	110
10	25	100
11	28	95
12	30	85

(iv). Draw the regression line on the scatter diagram.

(v). Predict the fasting blood glucose level for a patient who is administered 16 units of insulin.

(vi). Find the value of the standard error of the estimates.

(vii). Compute the coefficient of determination and interpret the results.

Solution

(i). The scatter plot in Figure 11.10 shows the relationship between insulin administered (in units) and fasting blood glucose level (in mg/dL) based on data from twelve patients.

Observations:

- There appears to be a negative correlation between the amount of insulin administered and the fasting blood glucose level. As the amount of insulin increases, the blood glucose level tends to decrease.
- The data points are generally clustered along a line, suggesting a linear relationship.

DRAFT

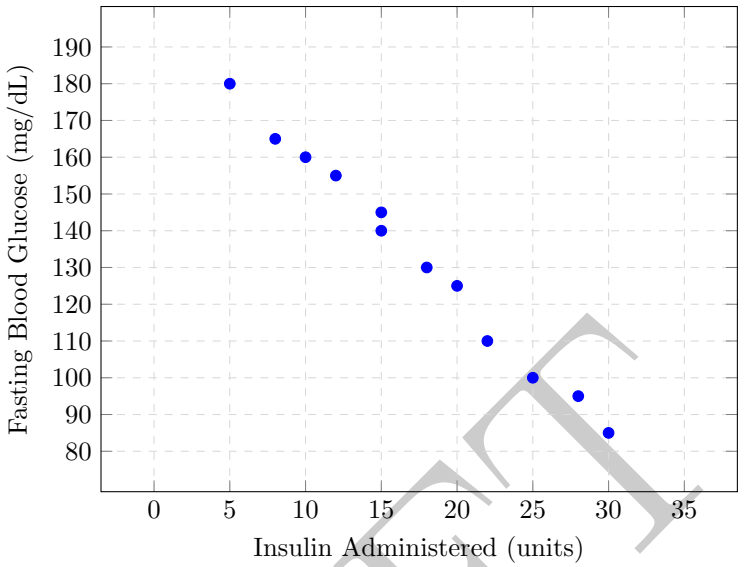


Figure 11.10: Scatter Plot of Insulin Administered vs. Fasting Blood Glucose

- The relationship is not perfectly linear, as there is some variability in glucose levels for patients who received the same amount of insulin (e.g., 15 units), which suggests other factors also influence blood glucose levels.

(ii). Simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad i = 1, 2, \ldots, 12 \tag{11.12}$$

where

- y_i = Fasting blood glucose level (mg/dL)
- x_i = Insulin administered (units)
- β_0 is the intercept, representing the expected blood glucose level when no insulin is administered.
- β_1 is the slope coefficient, representing the change in expected blood glucose level for each additional unit of insulin.
- e_i is the error term, assumed to be normally distributed with mean 0 and constant variance (σ^2).

Assumptions:

DRAFT

(i). **Linearity**: The relationship between insulin administered and blood glucose is linear.

(ii). **Independence of errors**: The blood glucose level of one patient does not influence the blood glucose level of another.

(iii). **Homoscedasticity**: The variance of the error term (σ^2) is constant for all levels of insulin administered.

(iv). **Normality of errors**: The error terms (e_i) are normally distributed.

(v). **Zero mean of errors**: The average value of the error term is zero, meaning the model is, on average, a correct representation of the relationship.

(iii). The estimated regression model is $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. The **ordinary least square (OLS) estimators** are given by:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \quad ; \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

The Least Squares Estimation First, we calculate the necessary sums and means from the data ($n = 12$):

x_i	y_i	x_i^2	$x_i y_i$
5	180	25	900
8	165	64	1320
10	160	100	1600
12	155	144	1860
15	140	225	2100
15	145	225	2175
18	130	324	2340
20	125	400	2500
22	110	484	2420
25	100	625	2500
28	95	784	2660
30	85	900	2550
Total = 208	**1590**	**4300**	**24925**

DRAFT

We have,

$$\bar{x} = \frac{208}{12} = 17, \qquad \bar{y} = \frac{1590}{12} = 132.5$$

Using the corrected results, we obtain

$$\widehat{\beta}_1 = -3.48$$

$$\widehat{\beta}_0 = 191.7$$

Thus, the estimated regression equation is

$$\widehat{y}_i = 191.7 - 3.48x_i.$$

The slope of the estimated regression equation ($\widehat{\beta}_1 = -3.48$) is negative, implying that as the amount of insulin administered increases, the fasting blood glucose level tends to decrease. Specifically, a one-unit increase in insulin administered is associated with an average decrease of approximately 3.48 mg/dL in the expected fasting blood glucose level.

(iv). The regression line along with the scatter plot is depicted in Figure 11.11.

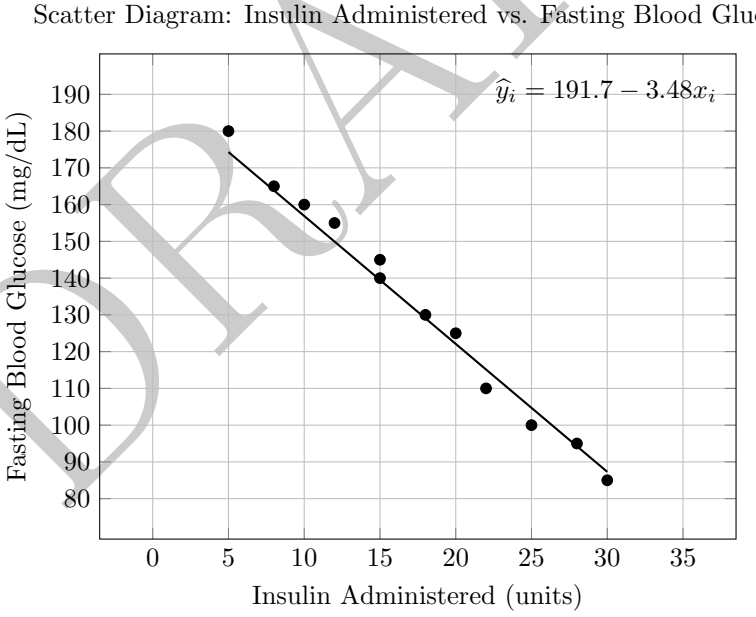


Figure 11.11: Scatter Plot of Insulin Administered vs. Fasting Blood Glucose with Regression Line

DRAFT

(v). To predict the fasting blood glucose level for a patient who is administered 16 units of insulin,

$$\widehat{y} = 191.7 - 3.48(16) = 191.7 - 55.68 = 136.02.$$

Hence, the predicted fasting blood glucose level is approximately 136.0 mg/dL.

x_i	y_i	$\widehat{y}_i = 191.7 - 3.48x_i$	$(y_i - \widehat{y}_i)^2$
5	180	174.30	32.49
8	165	163.86	1.30
10	160	156.90	9.61
12	155	149.94	25.60
15	140	139.50	0.25
15	145	139.50	30.25
18	130	129.06	0.88
20	125	122.10	8.41
22	110	115.14	26.42
25	100	104.70	22.09
28	95	94.26	0.55
30	85	87.30	5.29
Total			**95.83**

(vi). The Standard Error of the Estimates The estimated variance of the error term is

$$\widehat{\sigma}^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{95.83}{12-2} = 9.583, \qquad \widehat{\sigma} = \sqrt{9.583} \approx 3.09.$$

Hence, the standard error of the estimate is approximately 3.09 mg/dL.

(vii). Coefficient of Determination The coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST}.$$

Here, $SSE = 95.83$ and

$$SST = \sum(y_i - \bar{y})^2 = 10170, \qquad \bar{y} = 132.5.$$

Therefore,

$$R^2 = 1 - \frac{95.83}{10170} \approx 0.9906.$$

Interpretation: Approximately 99.06% of the total variability in fasting blood glucose levels is explained by the linear relationship with the amount of insulin administered. This indicates an excellent fit of the regression model.

DRAFT

Python Code Solution

Here is the Python code that performs the calculations and generates the plots for the example.

(i) and (iv). Scatter Plot and Regression Line This code uses the 'pandas', 'numpy', and 'matplotlib' libraries to load the data, calculate the regression line, and plot the results.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = {
  'insulin_administered': [5, 8, 10, 12, 15, 15, 18, 20, 22, 25, 28, 30],
  'fasting_blood_glucose': [180, 165, 160, 155, 140, 145, 130, 125, 110, 100, 95, 85]
}
df = pd.DataFrame(data)

# OLS estimators
n = len(df)
x = df['insulin_administered']
y = df['fasting_blood_glucose']
x_mean = np.mean(x)
y_mean = np.mean(y)
numerator = np.sum(x * y) - n * x_mean * y_mean
denominator = np.sum(x**2) - n * x_mean**2
beta_1_hat = numerator / denominator
beta_0_hat = y_mean - beta_1_hat * x_mean

# Plotting the regression line
plt.figure(figsize=(10, 6))
plt.scatter(x, y, label='Data Points')
plt.plot(x, beta_0_hat + beta_1_hat * x, color='red', label='Regression Line')
plt.title('Scatter Plot with Regression Line')
plt.xlabel('Insulin Administered (units)')
plt.ylabel('Fasting Blood Glucose (mg/dL)')
plt.legend()
plt.grid(True)
plt.show()

```

Listing 11.1: Python code to generate a scatter plot with the regression line.

(v). Prediction This code snippet predicts the blood glucose level for a new value of insulin using the calculated regression equation.

```
# Using the calculated coefficients
beta_0_hat = 198.25
beta_1_hat = -3.79
x_predict = 16
y_predict = beta_0_hat + beta_1_hat * x_predict
print(f"Predicted fasting blood glucose for {x_predict} units of insulin: {y_predict:.2f} mg/dL")

```

Listing 11.2: Python code to predict a new value.

(vi) and (vii). Standard Error and Coefficient of Determination The following code calculates the standard error of the estimates and the coefficient of determination (R^2).

```
# Using the previously calculated data
n = 12
x = np.array([5, 8, 10, 12, 15, 15, 18, 20, 22, 25, 28, 30])
y = np.array([180, 165, 160, 155, 140, 145, 130, 125, 110, 100, 95, 85])
beta_0_hat = 198.25
beta_1_hat = -3.79

# Calculate SSE
y_hat = beta_0_hat + beta_1_hat * x
sse = np.sum((y - y_hat)**2)
print(f"Sum of Squared Errors (SSE): {sse:.2f}")

# Calculate Standard Error
sigma_hat_squared = sse / (n - 2)
sigma_hat = np.sqrt(sigma_hat_squared)
print(f"Standard Error of the Estimates (sigma_hat): {sigma_hat:.2f}")

# Calculate SST and R^2
y_mean = np.mean(y)
sst = np.sum((y - y_mean)**2)
r_squared = 1 - (sse / sst)
print(f"Coefficient of Determination (R^2): {r_squared:.4f}")

```

Listing 11.3: Python code to calculate SSE, SST, and R-squared.

11.8.12 Interval Estimation and Hypothesis Testing

Confidence Interval for β_0

The confidence interval for β_0 can be computed using the standard formula for linear regression parameter estimates. The formula for the confidence interval for β_0 is:

$$\widehat{\beta}_0 \pm t_{\frac{\alpha}{2},(n-2)} \cdot \text{se}(\widehat{\beta}_0)$$

where:

- $t_{\alpha/2}$ is the critical value of the t-distribution with $n-2$ degrees of freedom at a significance level of $\alpha/2$ (where α is typically 0.05 for a 95% confidence interval),
- the standard error of the estimator $\widehat{\beta}_0$ is

$$\text{se}(\widehat{\beta}_0) = \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right)}.$$

Confidence Interval for β_1

The confidence interval for β_1 can be computed using the standard formula for linear regression parameter estimates. The formula for the confidence interval for β_1 is:

$$\widehat{\beta}_1 \pm t_{\frac{\alpha}{2},(n-2)} \cdot \text{se}(\widehat{\beta}_1)$$

where:

- $t_{\alpha/2}$ is the critical value of the t-distribution with $n-2$ degrees of freedom at a significance level of $\alpha/2$ (where α is typically 0.05 for a 95% confidence interval),
- the standard error of the estimator $\widehat{\beta}_1$ is

$$\text{se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where,

$$\widehat{\sigma} = \sqrt{\frac{\sum\limits_{i}^{n}(y_i - \widehat{y}_i)^2}{n-2}}$$

is the estimated standard error of the residuals (or the square root of the mean squared error, often obtained from the regression output). Once you have these values, you can compute the confidence interval.

DRAFT

11.8.13 The t-tests in Simple Linear Regression Model

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad ; \quad 1, 2, \ldots, n$$

To test the **individual significance of the independent variable**, the null hypothesis is

$$H_0 : \beta_1 = 0$$

Under H_0, the test statistic is given by:

$$t = \frac{\widehat{\beta}_1 - 0}{\text{se}(\widehat{\beta}_1)}$$

$$\text{where, se}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ and } \widehat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(y_i - \widehat{y}_i)^2}{n-2}}$$

We reject H_0 if $|t|$ exceeds the critical value $t_{\text{critical}} = t_{\frac{\alpha}{2},(n-2)}$ from the t-distribution with $n-2$ degrees of freedom, where n is the sample size.

Decision Rule for t-test

Classical Approach

- Set the significance level α.
- Calculate the critical value t_{critical} from the t-distribution with appropriate degrees of freedom.
- Decision Rule for each $\widehat{\beta}_1$:
  - If $|t| > t_{\text{critical}} = t_{\frac{\alpha}{2},(n-2)}$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is statistically significant.
  - If $|t| \leq t_{\text{critical}}$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is not statistically significant.

p-value Approach

- Calculate the p-value associated with each calculated t-statistic.
- Decision Rule for each $\widehat{\beta}_1$:
  - If p-value $< \alpha$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_1$ is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the cor-

responding coefficient $\widehat{\beta}_1$ is not statistically significant.

Note that the p-value can be obtained by using the formula:

$$p\text{-value} = 2 \times \min(P(T < -|t|), P(T > |t|))$$

11.8.14 Confidence Interval for $E(Y|X = x)$

- The Confidence Interval for $E(Y|X = x^*)$ provides a range of values where we expect the mean response Y to lie for a given value of x^* with a certain level of confidence.

- It is computed as:

$$\widehat{y}^* \pm t_{\frac{\alpha}{2},(n-2)} \cdot \text{se}(\widehat{y}^*)$$

  where $\widehat{y}^*$ is the predicted value of Y for a given x^*, $t_{\alpha/2}$ is the critical value of the t-distribution, n is the number of observations, and

$$\text{se}(\widehat{y}^*) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  is the standard error of the predicted value.

- x^* is the specific value of the independent variable for which you're predicting the response,

- The confidence level is typically chosen to be 95% ($\alpha = 0.05$).

For **Shorshe Ilish restaurant Sales Dataset** given in **Problem 11.7**, we have $\widehat{\sigma} = 13.829$. With $x^* = 10, \bar{x} = 14$, and $\sum(x_i - \bar{x})^2 = 568$, we have

$$\begin{aligned}
\text{se}(\widehat{y}^*) &= 13.829\sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} \\
&= 13.829\sqrt{0.1282} \\
&= 4.95
\end{aligned}$$

With $\widehat{y}^* = 110$ and a margin of error of $t_{0.025,8} \times \text{se}(\widehat{y}^*) = 2.306 \times 4.95 = 11.4147$, the 95% confidence interval for an average quarterly sales for the Shorshe Ilish restaurant located near campus for fixed $x^* = 10$ is

$$110 \pm 11.4147$$

where $t_{\frac{\alpha}{2},(n-1)} = t_{0.025,8} = 2.306$.

Thus, the 95% confidence interval for the mean quarterly sales, given a student population of 10,000, ranges from \$98585 to \$121415. Observe that the confidence interval for the mean value of Y will widen as x^* deviates further from x. This behavior is illustrated graphically in Figure 11.12.

DRAFT

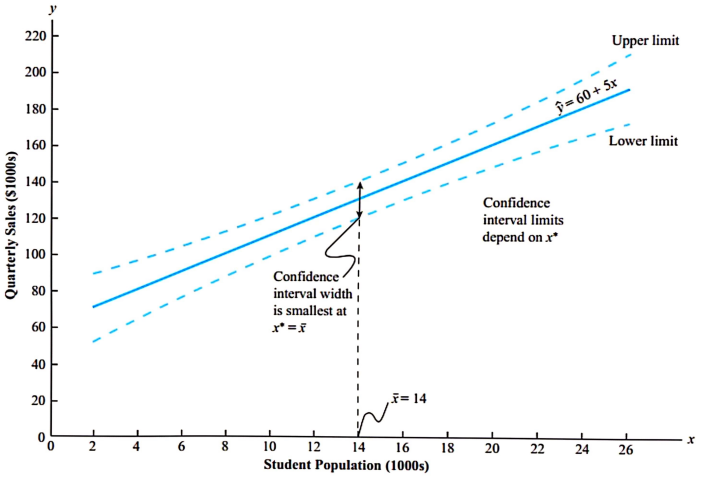


Figure 11.12: Confidence and prediction intervals for sales Y at given values of student population X

Prediction Interval for an Individual Value of Y

- The Prediction Interval for an Individual Value of Y provides a range of values where we expect a new observation of Y to lie with a certain level of confidence.
- It is wider than the Confidence Interval for the mean response $E(Y|X = x^*)$ because it accounts for the variability of individual observations around the regression line.
- It is computed as:

$$\widehat{y}^* \pm t_{\frac{\alpha}{2},(n-2)} \cdot s_{pred}$$

where

$$s^2_{pred} = \widehat{\sigma}^2 + \text{se}(\widehat{y}^*)^2 \text{ and hence } s_{pred} = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

is the standard error of the predicted value and $\widehat{y}^*$ is the predicted value of Y for a given x^*.

- The confidence level is typically chosen to be 95% ($\alpha = 0.05$).

For **Shorshe Ilish restaurant Sales Dataset** given in **Problem 11.7**, the estimated standard deviation corresponding to the prediction of quarterly sales for a new restaurant located need to the campus with 10,000 students, is

DRAFT

computed as follows

$$\begin{aligned} s_{pred} &= 13.829\sqrt{1+\frac{1}{10}+\frac{(10-14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

The 95% prediction interval for quarterly sales for the Shorshe Ilish restaurant located near campus can be found $t_{\frac{\alpha}{2},(n-1)} = 2.306$. Thus, with $\widehat{y}^* = 110$ and a margin of error of $t_{0.025} \times s_{pred} = 2.306 \times 14.69 = 33.875$, the 95% prediction interval is

$$110 \pm 33.875$$

In dollars, this prediction interval is \$76125 to \$143875.

Confidence intervals and prediction intervals become more precise as the value of the independent variable x^* approaches x. The typical shapes of confidence intervals and the broader prediction intervals are illustrated together in Figure 11.13.

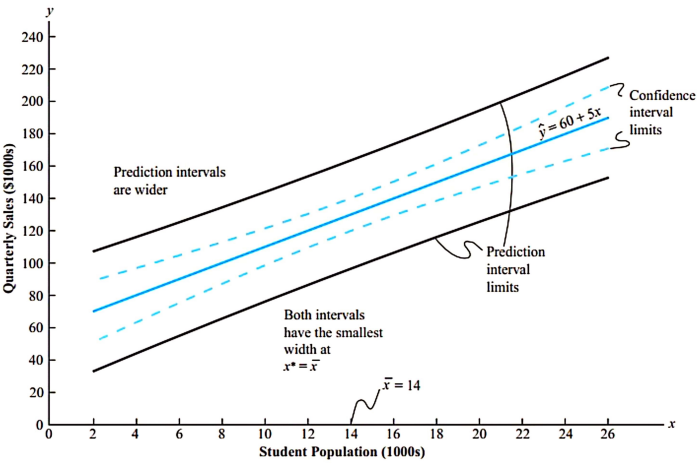


Figure 11.13: Confidence and prediction intervals for sales Y at given values of student population X

11.8.15 Python Code: Linear Regression Model

```
import numpy as np
import statsmodels.api as sm

# Define the data
```

```
    x = np.array([2, 6, 8, 8, 12, 16, 20, 20, 22, 26])
    y= np.array([58, 105, 88, 118, 117, 137, 157, 169, 149,
    202])

    # Add constant for intercept
    x_with_const = sm.add_constant(x)

    # Fit the model
    model = sm.OLS(y, x_with_const).fit()

    # Print summary of the regression model
    print(model.summary())

```

11.8.16 Exercises

1. What is meant by regression analysis? Distinguish between correlation and regression analysis.

2. Define and describe the main types of regression analysis (e.g., simple linear regression, multiple linear regression). Provide a real-world example for each type and explain why that particular type of regression would be used.

3. List and explain the four key assumptions of the Classical Simple Linear Regression model. For each assumption:

   (a) State the assumption mathematically where appropriate

   (b) Explain what it means in practical terms

   (c) Describe what happens if the assumption is violated

   (d) Suggest a diagnostic method to check the assumption

4. Consider the regression model: $\widehat{Y} = 15.2 + 3.8X$, where Y represents monthly sales revenue (in \$1000s) and X represents advertising spend (in \$1000s).

   (a) Interpret the intercept coefficient.

   (b) Interpret the slope coefficient.

   (c) Predict the monthly sales revenue when advertising spend is \$5,000.

   (d) What is the predicted change in sales revenue if advertising spend increases by \$2,000?

5. Match the following scenarios with the appropriate type of regression analysis:

   (a) Predicting a student's final grade based on hours studied.

DRAFT

(b) Modeling the relationship between house prices and multiple features (e.g., square footage, number of bedrooms).

(c) Estimating the effect of different levels of advertising spend on sales with a non-linear relationship.

6. The following sample of observations were randomly selected.

No. of TV Commercials (x)	2	5	1	3	4	1	5	3	4	2
Total Sales(y)	50	57	41	54	54	38	63	48	59	46

(i). Find the linear relationship between number of TV Commercials and total sales.

(ii). Fit a model y on x.

(iii). Find the coefficient of determination. Interpret your findings.

(iv). Find the adjusted R^2. Interpret your findings.

(v). Predict (or forecast) total sales when $x = 5$.

7. Consider the following sample of production volumes and total cost data for a manufacturing operation.

Production Volume (units)	Total Cost ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000
550	5500
615	6000

(i). Use these data to develop an estimated regression equation that could be used to predict the total cost for a given production volume. Interpret the value of regression intercept and slop coefficients. The company's production schedule shows 500 units must be produced next month. Predict the total cost for this operation

DRAFT

(ii). Compute the coefficient of determination. What percentage of the variation in total cost can be explained by production volume?

8. Given a dataset with the following statistics:

   - Pearson correlation coefficient $r = 0.8$
   - Standard deviation of x, $s_x = 4$
   - Standard deviation of y, $s_y = 5$

   Compute the following:

   (a) The regression coefficient $\widehat{\beta}_{xy}$ for predicting y from x.

   (b) The regression coefficient $\widehat{\beta}_{yx}$ for predicting x from y.

9. Consider two regression coefficients:

   - $\widehat{\beta}_{xy} = 1.5$
   - $\widehat{\beta}_{yx} = 0.7$

   Calculate the geometric mean of these regression coefficients and verify if it equals the absolute value of the Pearson correlation coefficient r.

10. You are given:

    - $\widehat{\beta}_{xy} = 2.0$
    - $\widehat{\beta}_{yx} = 0.5$
    - Standard deviation of x, $s_x = 3$
    - Standard deviation of y, $s_y = 6$

    Determine the Pearson correlation coefficient r using the given regression coefficients and standard deviations.

11. If $\widehat{\beta}_{xy} = 1.2$ and $\widehat{\beta}_{yx} = 0.9$, find the value of r^2 and compare it with the product of the regression coefficients.

12. For a simple linear regression model, you have the following sums of squares:

    - Total sum of squares (SST) = 300
    - Regression sum of squares (SSR) = 180

    Calculate the residual sum of squares (SS_{res}) and discuss its relation to the total and regression sum of squares.

DRAFT

13. Suppose the Pearson correlation coefficient between two variables is -0.6. If the standard deviations are $s_x = 5$ and $s_y = 10$, calculate:

    (a) $\widehat{\beta}_{xy}$

    (b) $\widehat{\beta}_{yx}$

    Discuss how a negative correlation affects the regression coefficients compared to a positive correlation.

14. You have the regression coefficients:

    - $\widehat{\beta}_{xy} = -0.4$
    - $\widehat{\beta}_{yx} = -1.1$

    Calculate the geometric mean of these coefficients and confirm if it matches the absolute value of the Pearson correlation coefficient.

15. Given the following dataset, perform a simple linear regression analysis:

Hours Studied	Exam Score
1	55
2	60
3	65
4	70
5	75

    (a) Compute the regression line equation $\widehat{Y} = \beta_0 + \beta_1 X$.

    (b) Plot the data and the regression line.

16. Discuss the key assumptions of the Classical Simple Linear Regression (CSLR) model. For each assumption, provide an example of a potential violation and explain how it could affect the results of the regression analysis.

17. Using the dataset from Exercise 15 , perform the following:

    (a) Calculate the OLS estimates for β_0 and β_1.

    (b) Derive the formula for the OLS estimator and apply it to find the estimates.

18. Calculate the estimated error variance and the standard error of the estimate for the dataset used in Exercise 15. Show all steps and formulas used in your calculations.

DRAFT

19. For the regression analysis in Exercise 15, compute the coefficient of determination (R^2). Interpret the value of R^2 in the context of the given data.

20. Explain the relationship between the coefficient of determination R^2 and the correlation coefficient r. If r is 0.8, what is R^2 and what does it signify about the regression model?

21. List and discuss the advantages and disadvantages of using R^2 as a measure of goodness-of-fit in regression analysis.

22. Given a multiple regression model with 3 predictors, calculate the adjusted R^2 if the R^2 is 0.85, the sample size is 50, and the number of predictors is 3. Interpret the result.

23. (a) Explain why R^2 always increases when adding variables to a regression model.
    (b) How does adjusted R^2 address this issue?
    (c) Under what circumstances might adjusted R^2 decrease when adding a variable?
    (d) In simple linear regression, what is the relationship between R^2 and the correlation coefficient r?

24. Given the regression output where $\widehat{y} = 1 + 2x$, construct a 95% confidence interval for β_1 (slope) if the standard error of β_1 is 0.5 and the sample size 28. Also, perform a hypothesis test to determine if β_1 is significantly different from 0.

25. For the regression analysis provided in Exercise 15, perform an ANOVA F-test to determine if the overall regression model is significant. Explain your decision rule and interpret the result.

26. Perform a hypothesis test to determine whether β_1 is significantly different from 0 in a regression model where the estimated β_1 is 2 and the standard error of β_1 is 0.4. Use a significance level of 0.05.

27. Calculate and interpret the confidence interval for the expected value $E(Y|X = x)$ using the dataset from Exercise 15.

28. A marketing analyst believes that social media advertising spend (X, in \$1000s) has a positive linear relationship with brand awareness score (Y, scale 1-100). From a sample of 15 campaigns, they obtain:

$$\widehat{Y} = 42.6 + 2.8X$$

with $se(\widehat{\beta}_1) = 0.6$, $s = 8.2$, and $R^2 = 0.73$.

    (a) Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$ at $\alpha = 0.05$ using a t-test.

DRAFT

(b) Calculate the 95% confidence interval for β_1.

(c) Test the overall significance of the regression model using an F-test at $\alpha = 0.05$.

(d) What can you conclude about the relationship between advertising spend and brand awareness?

29. An e-commerce company has the regression model:

$$\widehat{Y} = 125 + 15X$$

where Y is daily sales (\$) and X is number of website visitors (in hundreds). Given: $n = 20$, $s = 45$, $\bar{X} = 8$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 160$.

(a) Calculate the point prediction for daily sales when there are 600 website visitors.

(b) Calculate the 95% confidence interval for the mean daily sales $E(Y|X = 6)$.

(c) Calculate the 95% prediction interval for a single day's sales when $X = 6$.

(d) Explain the difference between the confidence interval and prediction interval.

(e) Which interval is wider and why?

30. Consider a dataset from a clinical study investigating the relationship between patient age (X) and systolic blood pressure (Y) in mmHg. You are given the following simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Table 11.9: Summary results of the regression model

	Parameter Estimates	
Parameter	Estimate	Standard Error
Intercept	95.4	3.2
Age (years)	0.85	0.12
R^2	0.42	

DRAFT

- Interpret each of the above results in the clinical context. What do they reveal about the relationship between age and systolic blood pressure?
- Does age play a statistically significant role in predicting systolic blood pressure? Justify your answer using appropriate statistical test.
- Calculate the predicted systolic blood pressure ($\widehat{Y}$) for a 60-year-old patient using the given regression model.
- Find the 95% confidence interval for the slope β_1. It is given that the critical value $t_{0.025} = 2.05$. Interpret this interval in terms of the effect of aging on blood pressure.
- Discuss the clinical implications of the R^2 value. What does this suggest about other factors that might influence blood pressure?
- Can you identify any potential clinical issues or statistical assumptions that should be considered when interpreting this regression output for patient care decisions?

31. A pharmaceutical researcher is studying the relationship between drug dosage (X) in mg and plasma drug concentration (Y) in mg/mL. The following simple linear regression model was fitted:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Table 11.10: Regression analysis results

	Parameter Estimates	
Parameter	Estimate	Standard Error
Intercept	2.1	0.75
Dosage (mg)	1.45	0.08
R^2	0.89	

- Interpret the regression coefficients in the pharmacokinetic context. What do they tell us about drug absorption and bioavailability?
- Test whether dosage significantly affects plasma concentration at $\alpha = 0.05$ level. State your hypothesis and conclusion.

DRAFT

- Predict the plasma concentration for a patient receiving a 25 mg dose of the drug.
- Construct a 99% confidence interval for the slope parameter, given that $t_{0.005} = 2.78$. What does this interval suggest about the dose-response relationship?
- Evaluate the goodness of fit using the R^2 value. Is this model suitable for clinical dose determination?
- What are the key assumptions of linear regression that must be validated before using this model for clinical dosing decisions? Mention at least three assumptions.

32. An epidemiologist is examining the association between Body Mass Index (X) and total cholesterol level (Y) in mg/dL in a population-based study. The simple linear regression results are:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Table 11.11: Cholesterol-BMI regression results

	Parameter Estimates	
Parameter	Estimate	Standard Error
Intercept	120.5	12.8
BMI (kg/m^2)	2.35	0.45
R^2	0.28	

- Interpret the regression parameters in the context of cardiovascular risk assessment. What do these estimates suggest about the BMI-cholesterol relationship?
- Is BMI a significant predictor of cholesterol levels? Perform the appropriate hypothesis test and state your conclusion at $\alpha = 0.01$ level.
- Estimate the total cholesterol level for an individual with BMI = 28.5 kg/m^2.
- Calculate the 95% confidence interval for the slope coefficient using $t_{0.025} = 1.98$. How would you communicate this finding to public health officials?

DRAFT

- Comment on the R^2 value from an epidemiological perspective. What does this imply about the complexity of cholesterol regulation?
- What are the limitations of this simple regression model for population health screening? Discuss confounding variables and other methodological considerations.

11.9 Multiple Linear Regression

In a multiple linear regression model, we extend the simple linear regression model (11.6) to include multiple independent variables. The general form of a multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i \tag{11.13}$$

where,

- y_i is the dependent variable (response variable) for the ith observation.
- $x_{1i}, x_{2i}, \ldots, x_{pi}$ are the independent variables for the ith observation.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients corresponding to the independent variables $x_{1i}, x_{2i}, \ldots, x_{pi}$.
- e_i is the error term or residual for the ith observation.

11.9.1 Model Assumptions

Under the multiple regression model (11.13), the following assumptions are made.

Assumptions

1. **Linearity**: The relationship between the dependent variable y_i and the independent variables $x_{1i}, x_{2i}, \ldots, x_{pi}$ is linear.
2. **Independence of Errors**: The error terms e_i are independent of each other.
3. **Homoscedasticity**: The variance of the error terms e_i is constant for all values of the independent variables.
4. **Normality of Errors**: The error terms e_i are normally distributed with mean zero.

DRAFT

5. **No Perfect Multicollinearity**: There is no perfect linear relationship among the independent variables.
6. **No Autocorrelation**: The errors (e) are not correlated with each other over time or across observations.

These assumptions are essential for valid estimation and interpretation of the classical regression model.

11.9.2 Estimation Procedure

To estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ in the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$

using ordinary least squares (OLS) in matrix notation, we define

$$\vec{\boldsymbol{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix} \vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \vec{\boldsymbol{e}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

In this case, the model can be written as

$$\vec{\boldsymbol{Y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{e}}$$

To find the estimated coefficients $\vec{\hat{\boldsymbol{\beta}}}$ of $\vec{\boldsymbol{\beta}}$ using ordinary least squares (OLS), we minimize the sum of squared errors. The vector of errors is a define as

$$\vec{\boldsymbol{e}} = \vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\boldsymbol{\beta}}$$

Calculation of Estimated Coefficients

The sum of squared residuals is given by:

$$\text{SSE} = \vec{\boldsymbol{e}}^T\vec{\boldsymbol{e}} = \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\boldsymbol{\beta}}\right)^T \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\boldsymbol{\beta}}\right)$$

To minimize SSE, we take the derivative with respect to $\vec{\boldsymbol{\beta}}$ and set it to zero:

$$\frac{\partial \text{SSE}}{\partial \vec{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\boldsymbol{\beta}}) = 0$$

Solving for $\vec{\boldsymbol{\beta}}$, we get: $\vec{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{\boldsymbol{Y}}$. Hence, the estimator of $\vec{\boldsymbol{\beta}}$ is

$$\vec{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{\boldsymbol{Y}}$$

This equation provides the estimated coefficients $\vec{\hat{\boldsymbol{\beta}}}$ that minimize the sum of squared residuals.

DRAFT

11.9.3 Estimation Procedure of Error Variance

To estimate the error variance $\widehat{\sigma}^2$, we use the following formula:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\widehat{\boldsymbol{\beta}}}\right)^T \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\widehat{\boldsymbol{\beta}}}\right)$$

where:

- $\vec{\widehat{\boldsymbol{\beta}}}$ is the vector of estimated coefficients obtained using ordinary least squares (OLS).
- n is the number of observations.
- p is the number of independent variables (excluding the intercept).

The term $\left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\widehat{\boldsymbol{\beta}}}\right)^T \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\widehat{\boldsymbol{\beta}}}\right)$ represents the sum of squared residuals, which measures the unexplained variability in the dependent variable after accounting for the effects of the independent variables. Dividing by $n-p-1$, the degrees of freedom for the error term, provides an unbiased estimate of the error variance $\widehat{\sigma}^2$.

11.9.4 Mean of the OLS Estimator

To find the mean of $\vec{\widehat{\boldsymbol{\beta}}}$:

$$\begin{aligned}
\vec{\widehat{\boldsymbol{\beta}}} &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{\boldsymbol{Y}} \\
&= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal (\boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{e}}) \\
&= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{X}\vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{\boldsymbol{e}} \\
&= \vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{\boldsymbol{e}}
\end{aligned}$$

Taking the expectation:

$$\begin{aligned}
E[\vec{\widehat{\boldsymbol{\beta}}}] &= E\left[\vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{\boldsymbol{e}}\right] = \vec{\boldsymbol{\beta}} + E\left[(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{\boldsymbol{e}}\right] \\
&= \vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal E[\vec{\boldsymbol{e}}] \\
&= \vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \vec{0} \\
&= \vec{\boldsymbol{\beta}}
\end{aligned}$$

So, the mean of the OLS estimator is:

$$E[\vec{\widehat{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}}$$

DRAFT

11.9.5 Variance of the OLS Estimator

To find the variance of $\widehat{\vec{\boldsymbol{\beta}}}$:

$$\widehat{\vec{\boldsymbol{\beta}}} = \vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \vec{\boldsymbol{e}}$$

The covariance matrix of $\widehat{\vec{\boldsymbol{\beta}}}$ is given by:

$$\begin{aligned} \text{Cov}(\widehat{\vec{\boldsymbol{\beta}}}) &= \text{Var}(\widehat{\vec{\boldsymbol{\beta}}}) \\ &= \text{Var}\left(\vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \vec{\boldsymbol{e}}\right) \\ &= \text{Var}\left((\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \vec{\boldsymbol{e}}\right) \end{aligned}$$

Since $\vec{\boldsymbol{e}} \sim N(\vec{0}, \sigma^2 \mathbf{I})$:

$$\text{Var}(\vec{\boldsymbol{e}}) = \sigma^2 \mathbf{I}$$

Using the properties of variance:

$$\begin{aligned} \text{Var}\left((\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \vec{\boldsymbol{e}}\right) &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \text{Var}(\vec{\boldsymbol{e}}) \boldsymbol{X} (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal (\sigma^2 \mathbf{I}) \boldsymbol{X} (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{X} (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \end{aligned}$$

So, the variance of the OLS estimator is:

$$\text{Var}(\widehat{\vec{\boldsymbol{\beta}}}) = \sigma^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

- The mean of the OLS estimator is:

$$E[\widehat{\vec{\boldsymbol{\beta}}}] = \vec{\boldsymbol{\beta}}$$

- The variance of the OLS estimator is:

$$\text{Var}(\widehat{\vec{\boldsymbol{\beta}}}) = \sigma^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

11.9.6 Coefficient of Determination

The coefficient of determination (R^2) measures the proportion of the variance in the dependent variable (Y) that is explained by the independent variables ($x_1, x_2, \ldots, x_p$) in the regression model.

$$R^2 = 1 - \frac{SSE}{SST}$$

where:

DRAFT

- $SSE = \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\hat{\boldsymbol{\beta}}}\right)^T \left(\vec{\boldsymbol{Y}} - \boldsymbol{X}\vec{\hat{\boldsymbol{\beta}}}\right)$ is the sum of squared errors (residuals), representing the unexplained variability in the dependent variable.
- $SST = (\vec{\boldsymbol{Y}} - \bar{\mathbf{Y}})^T(\vec{\boldsymbol{Y}} - \bar{\mathbf{Y}})$ is the total sum of squares, representing the total variability in the dependent variable.

Interpretation:

- R^2 ranges from 0 to 1, where a higher value indicates a better fit of the regression model to the data.
- R^2 represents the proportion of the variance in the dependent variable that is explained by the independent variables.
- For example, if $R^2 = 0.75$, it means that 75

11.9.7 Adjusted R^2

- the adjusted R^2 is denoted by $\bar{R}^2$ and is defined as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the total number of independent variables in the model (not including the constant term), and n is the sample size

- it can also be written as:

$$\bar{R}^2 = 1 - \frac{SSE/\mathrm{df}_e}{SST/\mathrm{df}_t}$$

where df_t is the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance

Hence,

$$\bar{R}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- the explanation of this statistic is almost the same as R^2 but it **penalizes the statistic as extra variables** are included in the model
- the term $\frac{n-1}{n-p-1}$ is called the **penalty** of using the more covariates in a model
- when the number of covariates p, increases, $(1 - R^2)$ will decrease, but $\frac{n-1}{n-p-1}$ will increase

DRAFT

- whether more covariates improve the explanatory power of a model depends on the **trade-of** between R^2 and the **penalty**

Interpretation:

- $\bar{R}^2$ penalizes the addition of unnecessary predictors to the model, unlike R^2.
- It is always less than or equal to R^2, and it increases only if the new term improves the model more than would be expected by chance.
- Therefore, $\bar{R}^2$ is often preferred for comparing the goodness of fit of models with different numbers of predictors.

11.9.8 Example Dataset and Regression Calculations

Consider a dataset with 10 observations and two independent variables (x_1 and x_2).

i	x_{1i}	x_{2i}	y_i
1	9	16	10
1	13	14	12
1	11	10	14
1	11	8	16
1	14	11	18
1	15	17	20
1	16	9	22
1	20	16	24
1	15	12	26
1	15	12	28

To fit a multiple regression model, we use the least squares method to estimate the coefficients β_0, β_1, and β_2.

The model equation is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

We calculate the estimated coefficients $\widehat{\beta}$ using the formula:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

where X is the design matrix of independent variables, Y is the vector of the dependent variable, and $\widehat{\beta}$ is the vector of estimated coefficients.

DRAFT

Calculation of $\vec{\hat{\beta}}$

For the given dataset, we have: $\boldsymbol{X} = \begin{bmatrix} 1 & 9 & 16 \\ 1 & 13 & 14 \\ 1 & 11 & 10 \\ 1 & 11 & 8 \\ 1 & 14 & 11 \\ 1 & 15 & 17 \\ 1 & 16 & 9 \\ 1 & 20 & 16 \\ 1 & 15 & 12 \\ 1 & 15 & 12 \end{bmatrix}$; $\vec{\boldsymbol{Y}} = \begin{bmatrix} 10 \\ 12 \\ \vdots \\ 28 \end{bmatrix}$.

Now, let's calculate $\boldsymbol{X}^T\boldsymbol{X}$, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $\boldsymbol{X}^T\vec{\boldsymbol{Y}}$, and $\vec{\hat{\boldsymbol{\beta}}}$.

Calculation of $\boldsymbol{X}^T\boldsymbol{X}$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 9 & 13 & 11 & \dots & 15 \\ 16 & 14 & 10 & \dots & 12 \end{bmatrix} \begin{bmatrix} 1 & 9 & 16 \\ 1 & 13 & 14 \\ 1 & 11 & 10 \\ \vdots & \vdots & \vdots \\ 1 & 15 & 12 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 139 & 125 \\ 139 & 2019 & 1757 \\ 125 & 1757 & 1651 \end{bmatrix}$$

Calculation of $\vec{\hat{\beta}}$

Using matrix inversion, we find:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{\det(\boldsymbol{X}^T\boldsymbol{X})}\text{adj}(\boldsymbol{X}^T\boldsymbol{X}) = \begin{bmatrix} 3.369 & -0.135 & -0.112 \\ -0.135 & 0.012 & -0.003 \\ -0.112 & -0.003 & 0.012 \end{bmatrix}$$

Now, we calculate

$$\vec{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{\boldsymbol{y}} = \begin{bmatrix} 2.821 \\ 1.591 \\ -0.475 \end{bmatrix}$$

DRAFT

Calculation of Error Variance Estimation

i	x_{1i}	x_{2i}	y_i	$\widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
1	9	16	10	9.542	0.209764
1	13	14	12	16.856	23.58074
1	11	10	14	15.573	2.474329
1	11	8	16	16.523	0.273529
1	14	11	18	19.871	3.500641
1	15	17	20	18.613	1.923769
1	16	9	22	24.003	4.012009
1	20	16	24	27.043	9.259849
1	15	12	26	20.988	25.12014
1	15	12	28	20.988	49.16814
Total	139	125	190	190	119.5229

- The error variance estimation, $\widehat{\sigma}^2$, is calculated as:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{119.5229}{10-2-1} = 17.0747$$

  where n is the number of observations, p is the number of predictors, y_i is the observed value, and $\widehat{y}_i$ is the predicted value.

- Coefficient of determination:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{119.522}{330} = 0.6378$$
$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} = 1 - \frac{(1-0.6378)(10-1)}{10-2-1} = 0.5343$$

Goodness-of-fit R^2 and Adjusted R^2

- The coefficient of determination $R^2 = 0.6378$ suggests that approximately 63.78% of the variance in the dependent variable (y) can be explained by the independent variables x_1 and x_2 included in the model.

- The Adjusted R^2 value takes into account the number of predictors and the sample size, providing a more conservative estimate of the model's goodness-of-fit. In this case, $\bar{R}^2 = 0.5343$ indicating that approximately 53.43% of the variance in the dependent variable(y) is explained by the independent variables x_1 and x_2 after adjusting for the number of predictors and the sample size.

DRAFT

11.9.9 F-test in Multiple Regression

- The F-test in multiple regression assesses the overall significance of the regression model.
- It tests whether at least one of the independent variables has a non-zero coefficient.
- The null hypothesis H_0 for the F-test is:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the independent variables.

11.9.10 ANOVA Table in Regression Analysis

- The ANOVA table in multiple regression assesses the overall significance of the regression model.
- It partitions the total variance in the dependent variable into explained variance and unexplained variance.
- The table includes sums of squares (SS), degrees of freedom (df), mean squares (MS), and the F-test statistic.

Source of Variation	**SS**	**df**	**MS**	**F**
Regression	SSR	p	$MSR = \frac{SSR}{p}$	
Residual (Error)	SSE	$n - p - 1$	$MSE = \frac{SSE}{n-p-1}$	$F = \frac{MSR}{MSE}$
Total	SST	$n - 1$		

- Reject the null hypothesis H_0 if the calculated F-statistic is greater than the critical value from the F-distribution.

Decision Rule for ANOVA F-test

Classical Approach

- Set the significance level α.
- Calculate the critical value F_{critical} from the F-distribution with appropriate degrees of freedom.
- Decision Rule:
  - If calculated $F > F_{\text{critical}}$, reject H_0 and conclude that the regression model is statistically significant.

DRAFT

- If calculated $F \leq F_{\text{critical}}$, fail to reject H_0 and conclude that the regression model is not statistically significant.

p-value Approach

- Calculate the p-value associated with the calculated F-statistic.
- Decision Rule:
  - If p-value $< \alpha$, reject H_0 and conclude that the regression model is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the regression model is not statistically significant.

Example: ANOVA Table

	df	*SS*	*MS*	*F*	*p-value*
Regression	2	210.4651	105.2326	6.1625	0.0286
Residual	7	119.5349	17.0764		
Total	9	330			

11.9.11 The t-tests in Multiple Regression

- The t-tests in multiple regression assess the significance of individual coefficients (parameters) in the model.
- Each t-test tests the null hypothesis that the corresponding coefficient is zero.
- The t-test statistic for the coefficient $\widehat{\beta}_i$ is calculated as:

$$t = \frac{\widehat{\beta}_i}{\text{se}(\widehat{\beta}_i)}$$

where $\widehat{\beta}_i$ is the estimated coefficient, and $\text{se}(\widehat{\beta}_i)$ is its standard error.

Decision Rule for t-tests in Multiple Regression

Classical Approach

- Set the significance level α.
- Calculate the critical value t_{critical} from the t-distribution with appropriate degrees of freedom.

DRAFT

- Decision Rule for each $\widehat{\beta}_i$:
  - If $|t| > t_{\text{critical}}$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is statistically significant.
  - If $|t| \leq t_{\text{critical}}$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is not statistically significant.

p-value Approach

- Calculate the p-value associated with each calculated t-statistic.
- Decision Rule for each $\widehat{\beta}_i$:
  - If p-value $< \alpha$, reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is statistically significant.
  - If p-value $\geq \alpha$, fail to reject H_0 and conclude that the corresponding coefficient $\widehat{\beta}_i$ is not statistically significant.

Problem 11.10. *Consider a dataset containing information about the quarterly sales of a restaurant chain, the student population near the restaurant, and the advertising budget. The following data has been collected:*

Restaurant (1000s)	***Student Population (1000s of Dollars)***	***Advertising Budget (1000s of Dollars)***	***Quarterly Sales*** *(y)*
1	2	5	58
2	6	12	105
3	8	10	88
4	8	15	118
5	12	20	117
6	16	22	137
7	20	25	157
8	20	26	169
9	22	24	149
10	26	30	202

(i) Using the above data, formulate a multiple linear regression model where the quarterly sales (y) is the dependent variable, and both the student population (x_1) and the advertising budget (x_2) are independent variables.

(ii) Write down the assumptions of the multiple linear regression model.

DRAFT

(iii) Estimate the regression coefficients for the model formulated in Question 1 using the least squares method. Interpret the coefficients for both the student population and advertising budget.

(iv) Based on the regression model obtained in part (a), what would be the estimated quarterly sales for a restaurant located near a campus with 16,000 students and an advertising budget of $20,000?

(v) Calculate the standard error of the estimates for the model obtained in Question (iii).

(vi) How would you evaluate the goodness-of-fit for the regression model? What statistical metrics would you consider, and why?

(vii) Create scatter plots to show the relationship between:

- *Student population and quarterly sales.*
- *Advertising budget and quarterly sales.*

Overlay the regression lines on these plots and comment on the observed relationships.

(viii) If a new restaurant is to be established near a campus with a student population of 18,000 and an advertising budget of $25,000, use the regression model from Question 2 to predict the quarterly sales.

(ix) Discuss the limitations of using the linear regression model in this context. What are some factors not considered by the model that could affect the accuracy of your predictions?

(x) Test the significance of each regression coefficient at the 5% significance level. Clearly state the null and alternative hypotheses, the test statistic, and your conclusion.

(xi) Construct a 95% confidence interval for the coefficients of the student population and advertising budget. Interpret the intervals.

DRAFT

Solution

(i).

The multiple linear regression model for this example is formulated as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

- y_i represents the **quarterly sales** (in thousands of dollars).
- x_{1i} represents the **student population** (in thousands).
- x_{2i} represents the **advertising budget** (in thousands of dollars).

- β_0 is the **intercept**, the expected sales when both predictors are zero.
- β_1 is the coefficient for the student population, representing the expected change in sales for a one-unit increase in student population, holding the advertising budget constant.
- β_2 is the coefficient for the advertising budget, representing the expected change in sales for a one-unit increase in advertising budget, holding the student population constant.
- e_i is the **error term**, which captures unobserved factors affecting sales.

(ii).

The assumptions of the multiple linear regression model are:

(i). **Linearity**: The relationship between the dependent variable (y) and the independent variables (x_1, x_2) is linear.

(ii). **Independence of errors**: The error terms (ϵ_i) are independent of each other.

(iii). **Homoscedasticity**: The variance of the error terms (σ^2) is constant across all levels of the independent variables.

(iv). **Normality of errors**: The error terms are normally distributed.

(v). **No multicollinearity**: The independent variables (x_1 and x_2) are not highly correlated with each other.

(iii).

After fitting the regression model to the data, the estimated coefficients are:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 \\ &= 38.62 + 0.97x_1 + 4.11x_2\end{aligned}$$

Interpretation:

- The intercept $\widehat{\beta}_0 = 38.62$ suggests that when both the student population and advertising budget are zero, the estimated quarterly sales would be \$38,620. However, this may not have a practical interpretation.
- The coefficient $\widehat{\beta}_1 = 0.97$ suggests that for each additional 1000 students, quarterly sales increase by \$970, holding the advertising budget constant.
- The coefficient $\widehat{\beta}_2 = 4.11$ suggests that for each additional \$1000 spent on advertising, quarterly sales increase by \$4110, holding the student population constant.

DRAFT

(iv).

Substituting $x_1 = 16$ and $x_2 = 20$ into the estimated regression equation:

$$\widehat{y} = 38.62 + 0.97(16) + 4.11(20) = 141.30$$

Therefore, the estimated quarterly sales would be \$141,300.

(v).

The standard error of the estimate is calculated as:

$$\text{Standard Error} = \sqrt{\frac{\sum(y_i - \widehat{y}_i)^2}{n - p}}$$

where n is the number of observations and p is the number of predictors including the intercept. From the regression model, the standard error of the estimate is calculated as:

$$\text{Standard Error} = 12.44$$

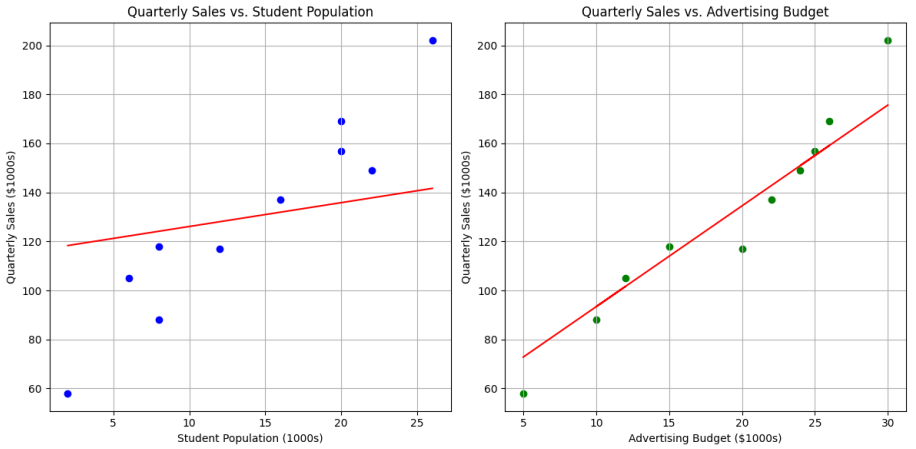
(vi).

The goodness-of-fit of the regression model can be evaluated using the following metrics:

- $R^2 = 0.944$: Represents the proportion of variance in the dependent variable that is explained by the independent variables. A higher R^2 indicates a better fit.
- Adjusted $R^2 = 0.928$: Adjusts the R^2 value based on the number of predictors, providing a more accurate measure in models with multiple predictors.
- F-statistic: $F = 59.36$, with a p-value of 0.00004, tests the overall significance of the regression model. A significant F-statistic indicates that the model explains a significant portion of the variance in the dependent variable.
- Residual plots: Used to check for homoscedasticity and the normality of the residuals.

DRAFT

(vii).



- The scatter plot of student population vs. quarterly sales shows a positive linear relationship, indicating that as the student population increases, quarterly sales also increase.
- The scatter plot of advertising budget vs. quarterly sales also shows a positive linear relationship, indicating that an increase in the advertising budget leads to higher quarterly sales.

(viii).

Substituting $x_1 = 18$ and $x_2 = 25$ into the estimated regression equation:

$$\widehat{y} = 38.62 + 0.97(18) + 4.11(25) = 162.28$$

The predicted quarterly sales would be \$162,280.

(ix).

Limitations of the linear regression model in this context include:

- The model assumes a linear relationship between the variables, which may not always hold true.
- It does not account for potential interactions between the student population and advertising budget.
- The model does not consider external factors like economic conditions, competition, or seasonal variations that could impact sales.
- The model assumes that the relationships are constant over time, which might not be the case in reality.

DRAFT

(x).

For each coefficient:

- **Null Hypothesis (H_0)**: The coefficient is equal to zero ($\beta_i = 0$).
- **Alternative Hypothesis (H_1)**: The coefficient is not equal to zero ($\beta_i \neq 0$).

Test Statistics (from the regression output):

- For the intercept (β_0): $t = 3.229$, $p = 0.014$.
- For student population (β_1): $t = 0.534$, $p = 0.610$.
- For advertising budget (β_2): $t = 2.287$, $p = 0.056$.

Conclusion:

- The intercept is statistically significant ($p < 0.05$).
- The coefficient for student population is not statistically significant ($p > 0.05$), so we fail to reject the null hypothesis.
- The coefficient for the advertising budget is marginally significant at the 5% level ($p \approx 0.056$), close to the threshold, so we might consider it as not significant or marginally significant depending on the context.

(xi).

Using the regression output, the 95% confidence intervals are:

- For student population (β_1): $-3.337 \leq \beta_1 \leq 5.283$
- For advertising budget (β_2): $-0.140 \leq \beta_2 \leq 8.369$

Interpretation:

- For student population: We are 95% confident that the true coefficient lies within the interval $[-3.337, 5.283]$. Since this interval includes zero, it suggests that the effect of student population on sales may not be significant.
- For advertising budget: We are 95% confident that the true coefficient lies within the interval $[-0.140, 8.369]$. Since this interval also includes zero, it suggests that the effect of advertising budget on sales may not be significant.

DRAFT

Python Code: Multiple Linear Regression Model

Consider the Problem 11.10. Suppose we have data on quarterly sales (y), student population (x_1), and advertising budget (x_2). The Python code to solve this problem is in the following.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Example Data: Student Population, Advertising Budget, and Quarterly Sales
data = {
    'Student Population (1000s)': [2, 6, 8, 8, 12, 16, 20, 20, 22, 26],
    'Advertising Budget ($1000s)': [5, 12, 10, 15, 20, 22, 25, 26, 24, 30],
    'Quarterly Sales ($1000s)': [58, 105, 88, 118, 117, 137, 157, 169, 149, 202]
}

# Convert the data to a DataFrame
df = pd.DataFrame(data)

# Define the independent variables (X) and dependent variable (y)
X = df[['Student Population (1000s)', 'Advertising Budget ($1000s)']]
y = df['Quarterly Sales ($1000s)']

# Add a constant to the independent variables matrix
X = sm.add_constant(X)

# Fit the multiple linear regression model
model = sm.OLS(y, X).fit()

# Print the summary of the regression model
print(model.summary())

# Plotting the relationship
# (Optional) You can create scatter plots to visualize the relationship between each independent variable and the dependent variable.

plt.figure(figsize=(12, 6))

# Scatter plot for Student Population vs Quarterly Sales
plt.subplot(1, 2, 1)
```

```
plt.scatter(df['Student Population (1000s)'], y, color='blue')
plt.plot(df['Student Population (1000s)'], model.params[0] + model.params[1] * df['Student Population (1000s)'] + model.params[2] * df['Advertising Budget ($1000s)'].mean(), color='red')
plt.title('Quarterly Sales vs. Student Population')
plt.xlabel('Student Population (1000s)')
plt.ylabel('Quarterly Sales ($1000s)')
plt.grid(True)

# Scatter plot for Advertising Budget vs Quarterly Sales
plt.subplot(1, 2, 2)
plt.scatter(df['Advertising Budget ($1000s)'], y, color='green')
plt.plot(df['Advertising Budget ($1000s)'], model.params[0] + model.params[1] * df['Student Population (1000s)'].mean() + model.params[2] * df['Advertising Budget ($1000s)'], color='red')
plt.title('Quarterly Sales vs. Advertising Budget')
plt.xlabel('Advertising Budget ($1000s)')
plt.ylabel('Quarterly Sales ($1000s)')
plt.grid(True)

plt.tight_layout()
plt.show()

# Prediction example: Predict quarterly sales for a restaurant with a student population of 16,000 and an advertising budget of $20,000
new_data = pd.DataFrame({'const': [1], 'Student Population (1000s)': [16], 'Advertising Budget ($1000s)': [20]})
predicted_sales = model.predict(new_data)[0]
print(f'Predicted Quarterly Sales: ${predicted_sales:.2f} (in $1000s)')

# Standard Error of the Estimates
standard_error = np.sqrt(model.mse_resid)
print(f'Standard Error of the Estimates: {standard_error:.2f}')

```

Figure 11.14: Python output for the practice example.

```
================================================================================
Dep. Variable:     Quarterly Sales ($1000s)   R-squared:                       0.944
Model:                                 OLS   Adj. R-squared:                  0.928
Method:                      Least Squares   F-statistic:                     59.36
Date:                     Mon, 12 Aug 2024   Prob (F-statistic):           4.07e-05
Time:                             06:28:51   Log-Likelihood:                -36.552
No. Observations:                       10   AIC:                             79.10
Df Residuals:                            7   BIC:                             80.01
Df Model:                                2
Covariance Type:                 nonrobust
================================================================================
                                 coef    std err   t      P>|t|    [0.025   0.975]
--------------------------------------------------------------------------------
const                         38.6235     11.962   3.229   0.014    10.338   66.909
Student Population (1000s)     0.9727      1.823   0.534   0.610    -3.337    5.283
Advertising Budget ($1000s)    4.1142      1.799   2.287   0.056    -0.140    8.369
==============================================================================
Omnibus:                        0.630   Durbin-Watson:                   2.261
Prob(Omnibus):                  0.730   Jarque-Bera (JB):                0.535
Skew:                           0.038   Prob(JB):                        0.765
Kurtosis:                       1.869   Cond. No.                         88.2
==============================================================================
```

11.9.12 Exercises

1. Explain the difference between simple linear regression and multiple linear regression. How does the addition of more predictors in multiple linear regression improve the model?

2. What assumptions must be met for multiple linear regression analysis to be valid? Explain why each assumption is important.

3. What is R^2 in the context of multiple linear regression? How is it different from adjusted R^2, and why might adjusted R^2 be preferred in some cases?

4. The following summary results are for a multiple regression model:

Table 11.12: Model Summary Statistics

Statistic	**Value**
R-squared	0.76
Adjusted R-squared	0.74
F-statistic	38.42
p-value (F-statistic)	0.0001

DRAFT

Table 11.13: Regression Coefficients

Variable	**Estimate**	**Standard Error**	**p-value**
Intercept	2.50	0.65	0.001
X_1	0.75	0.10	0.000
X_2	-0.55	0.15	0.102
X_2	1.20	0.20	0.000

(i). What is the regression model corresponding to the summary results given in Table 11.13? Describe the assumptions of multiple regression analysis. What diagnostic plots can be used to check these assumptions?

(ii). Explain the R-squared value given in Table 11.12, which indicates the model's performance. How does the adjusted R-squared value improve on this interpretation?

(iii). Given the model summary in Table 11.13, write the estimated model and

(a) Interpret the coefficient of X_1.

(b) What does the intercept term represent in this context?

(iv). Based on the p-values of the coefficients in Table 11.13, identify which predictors are statistically significant at the 5% significance level and explain why.

(v). Calculate the predicted value of Y when $X_1 = 4$, $X_2 = 2$, and $X_3 = 3$. Show all steps in your calculation.

(vi). The F-statistic and its corresponding p-value are given in the model summary. Explain what these values indicate about the overall model.

5. Describe how hypothesis testing is used to determine the significance of the regression coefficients in a multiple linear regression model. What is the null hypothesis for each coefficient, and how is it tested?

6. What is the purpose of the F-test in multiple linear regression? How does it differ from t-tests for individual coefficients?

11.10 Regression Model Diagnostics

Regression model diagnostics are critical for ensuring that the assumptions underlying a regression model are met. These diagnostics help identify issues such as non-linearity, heteroscedasticity, outliers, and multicollinearity, which can affect the validity and interpretability of the model.

DRAFT

11.10.1 Assumptions of Linear Regression

The linear regression model is based on several key assumptions that must be checked to ensure the model is appropriate for the data:

- **Linearity**: The relationship between the independent variables and the dependent variable should be linear.
- **Independence**: Observations should be independent of each other.
- **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables.
- **Normality of Residuals**: The residuals (errors) of the model should be approximately normally distributed.
- **No Multicollinearity**: Independent variables should not be highly correlated with each other.

11.10.2 Residual Plots

Residuals of the model given in Equation (11.6) are defined as

$$\widehat{e}_i = y_i - \widehat{y}_i \quad \text{for } i = 1, 2, \ldots, n.$$

These residuals can be regarded as the 'observed' errors. They are not the same as the unknown true errors

$$e_i = y_i - E(y_i) \quad \text{for } i = 1, 2, \ldots, n.$$

If the model is appropriate for the data at hand, the residuals should reflect the properties assumed for e_i (i.e., independence, normality, zero mean, and constant variance). For diagnostic purposes, we sometimes use the **semistudentized residual**, which is defined as

$$\widehat{e}_i^* = \frac{\widehat{e}_i - \bar{\widehat{e}}}{\sqrt{MSE}} = \frac{\widehat{e}_i}{\sqrt{MSE}}, \tag{11.14}$$

where

$$MSE = \sum_{i=1}^{n} \frac{(\widehat{e}_i - \bar{\widehat{e}})^2}{n-k} = \sum_{i=1}^{n} \frac{\widehat{e}_i^2}{n-k} = s^2$$

and $k = p + 1$.

If $\sqrt{MSE}$ were an estimate of the standard deviation of $\widehat{e}_i$, then $\widehat{e}_i^*$ would be the studentized residual. However, note that the standard deviation of e_i is not equal to $\sqrt{MSE}$; it varies for each $\widehat{e}_i$. The estimator $\sqrt{MSE}$ is only an approximation of the standard deviation of e_i; hence, we refer to $\widehat{e}_i^*$ as a

DRAFT

semistudentized residual.

Residual plots are a crucial tool for diagnosing issues in a regression model. A residual plot is a scatterplot of the residuals (errors) on the y-axis and the predicted values or one of the independent variables on the x-axis. The following plots of residuals (or semistudentized residuals) will be utilized here for this purpose: '

1. Plot of residuals against predictor variable.
2. Plot of absolute or squared residuals against predictor variable.
3. Plot of residuals against fitted values.
4. Plot of residuals against time or outlier sequence.
5. Plots of residuals against omitted predictor variables.
6. Box plot of residuals.
7. Normal probability plot of residuals.

A few questions to consider when you analyze plots

1. Do the residuals follow any pattern indicating nonlinearity?
2. Are there any outliers?
3. Does the assumption of constant variance look correct?
4. Label any qualitative variables on the plot. Any patterns?

- **Detecting Non-linearity**: If the residuals show a systematic pattern (e.g., curvature), this indicates that the relationship between the independent and dependent variables may not be linear.
- **Identifying Heteroscedasticity**: If the residuals display a funnel shape (widening or narrowing), this suggests heteroscedasticity, meaning the variance of the errors is not constant.

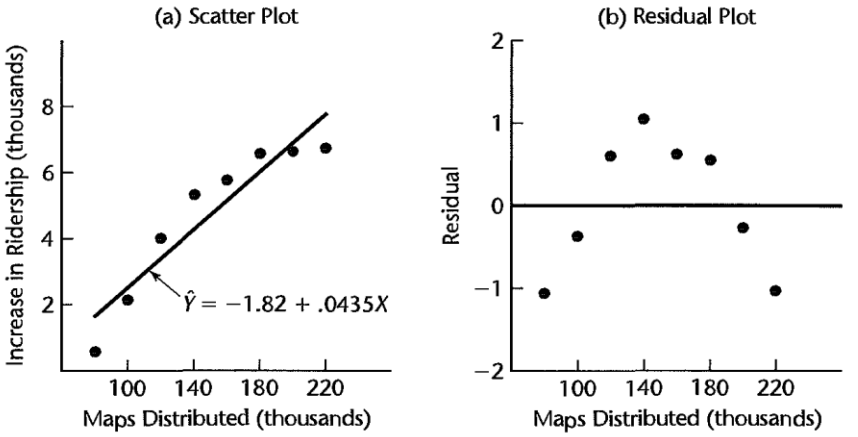
Univariate Plots of x and y

Univariate plots (See Figure 11.15(a)) of x and y are essential tools for exploring the individual distributions of the predictor variable x and the response variable y in a regression analysis. These plots, such as histograms, box plots, or density plots, allow us to visualize the central tendency, spread, and shape of each variable's distribution. By examining these plots, we can identify characteristics such as skewness, kurtosis, **outliers**, or gaps in the data, which may influence the relationship between x and y. Understanding the univariate distributions of x and y helps in choosing appropriate transformations or modeling approaches,

DRAFT

and provides a foundation for interpreting the results of more complex analyses, such as regression or correlation.

Figure 11.15: Scatter Plot and Residual Plot Understanding Nonlinear Regression Function

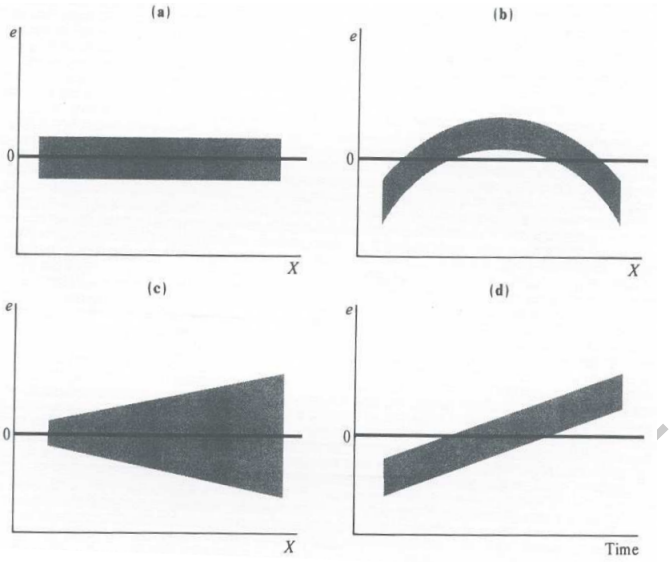


Plot of the Residuals versus x

Plot of the Residuals versus x (See Figure 11.15(b)) is more effective in detecting nonlinearity than a scatter plot (e.g., nonlinearity may be more prominent on a residual plot compared to on a scatter plot). It can also indicate other forms of model departure besides nonlinearity (e.g., non-constancy of variance). See Figures 11.16. For any observable pattern(s) in the plot of the residuals versus x indicates a problem with the model assumptions!

DRAFT

Figure 11.16: Prototype Residual Plots.



Plot of the Residuals versus $\widehat{y}$

For simple linear regression (with a single predictor), the residuals versus $\widehat{y}$ plot contains the same information as the plot of residuals versus X, but on a different scale. For multiple linear regression, this plot allows us to examine patterns in the residuals as y increases. Ideally, there should be no systematic patterns.

Plot of the Residuals versus Time

The plot of residuals versus time (See Figure 11.17), also known as a sequence plot, is only meaningful when data are collected in a time sequence or some other type of ordered sequence. In this plot, any discernible pattern suggests a lack of independence among the residuals, indicating potential issues with the model.

DRAFT

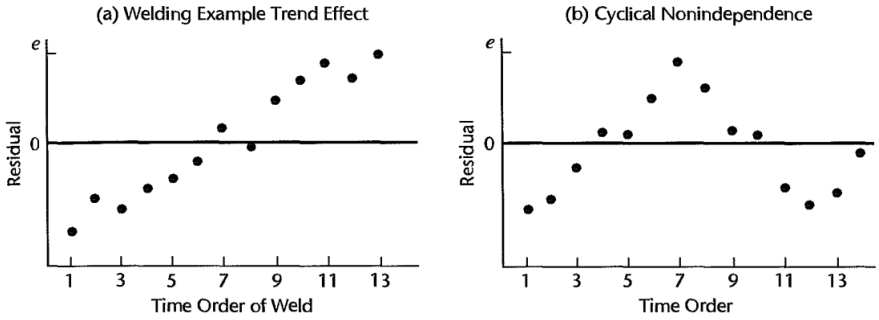


Figure 11.17: Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.

Plot of the Semistudentized Residuals versus X

Semistudentized residuals are defined in Equation (11.14). The plot shown in Figure 11.18 is particularly useful for identifying outliers in the data. By using the semistudentized version of the residual, it becomes easier to detect potential outliers, as the residuals are scaled relative to the variability of the data. Cases where $\widehat{e}_i^*$ falls outside the range $(-3, 3)$ can be considered outliers, indicating observations that deviate significantly from the fitted model.

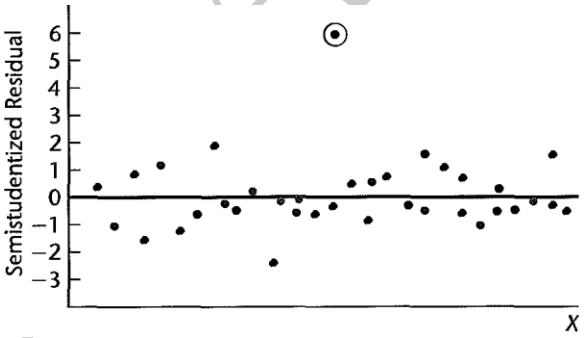


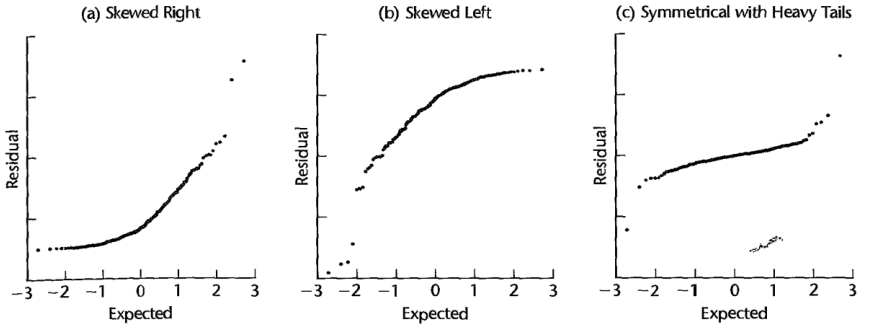
Figure 11.18: Residual Plot with Outlier.

Plot of the Residuals in a Normal Quantile Plot

A Q-Q plot, also known as a quantile-quantile (QQ) plot (See Figure 11.16), is used to assess the normality of the residuals in a regression model. In this plot, the ordered residuals are plotted against their corresponding quantiles from a normal distribution. If the residuals are normally distributed, the points should lie approximately along a straight line. Deviations from this line, especially in the tails of the distribution, indicate non-normality in the residuals. This can

DRAFT

Figure 11.19: Normal Probability Plots when Error Term Distribution Is Not Normal.



signal potential issues with the model, such as the presence of outliers, skewness, or other departures from the assumptions of normality.

Added-Variable plot

An Added-Variable plot, also known as a partial regression plot, is a diagnostic tool used in multiple linear regression to assess the relationship between a specific predictor variable and the response variable, while accounting for the influence of other predictors in the model. The Added-Variable plot is created by plotting the residuals from regressing the response variable on the other predictors against the residuals from regressing the specific predictor on the same set of predictors. This helps in identifying the unique contribution of the predictor of interest after removing the effects of other variables in the model.

In some cases, an Added-Variable plot may also involve plotting residuals against potential predictor variables that are not included in the model but could have important effects on the response. Any distinctive pattern(s) in such a plot may indicate that the model could be improved by adding the omitted predictor variable(s). If the plot shows a clear linear trend, it suggests that the predictor has a significant linear relationship with the response variable. However, deviations from linearity or a lack of pattern may indicate that the predictor does not contribute significantly to the model, or that a more complex, nonlinear relationship exists.

11.10.3 Formal Tests

Graphical analysis can often be subjective, especially when patterns are not very distinctive. In such cases, formal tests can be considered to assess potential violations of the model assumptions.

- Tests for Normality

DRAFT

- Test for Autocorrelation
- Tests for Non-Constancy of Variance
  - Breusch-Pagan Test
- Outlier Identification
- Lack-of-Fit Test

Tests for Normality

The goodness-of-fit tests can be used to examine the normality of the error terms

- Chi-Square test
- Shapiro-Wilk test,
- Kolmogorov-Smirnov test,
- Lilliefors test

Shapiro-Wilk test for Normality

The Shapiro-Wilk test is a statistical test used to assess the normality of a dataset. Specifically, it tests the null hypothesis that a sample comes from a normally distributed population. The test calculates a W statistic, which measures how well the data conform to a normal distribution. A W value close to 1 indicates that the data are approximately normally distributed, while a W value significantly less than 1 suggests a departure from normality. The associated p-value determines whether the null hypothesis can be rejected; a small p-value (typically less than 0.05) indicates that the data significantly deviate from normality. The Shapiro-Wilk test is particularly powerful for detecting departures from normality in small to moderately sized samples, making it a widely used tool in statistical analysis. Using a hypothetical dataset, the Python code for Shapiro-Wilk test is explained in the next section.

Example: Toluca Company dataset

The Toluca Company, which manufactures refrigeration equipment and replacement parts, aimed to determine the optimal lot size for a specific replacement part during a cost improvement program. Production involves setup, machining, and assembly, with labor hours varying by lot size. To establish the relationship between lot size and labor hours, data from 25 production runs were collected over a stable six-month period. This data, shown in Table 11.14, records lot sizes and corresponding work hours. Lot sizes are multiples of 10, as

DRAFT

per company policy. The conditions during this period were expected to remain consistent for the next three years.

Table 11.14: Toluca Company Example.

Lot Size	Work Hrs	Lot Size	Work Hrs
80	399	30	121
50	221	90	376
70	361	60	224
120	546	80	352
100	353	50	157
40	160	70	252
90	389	20	113
110	435	100	420
30	212	50	268
90	377	110	421
30	273	90	468
40	244	80	342
70	323		

We perform Shapiro-Wilk test for Normality using Python in the following.

Python Code: Shapiro-Wilk test for Normality

```
# Shapiro-Wilk test for Normality
import numpy as np
import statsmodels.api as sm
from scipy.stats import shapiro
# Define the data
LotSize = np.array([80, 30, 50, 90, 70, 60, 120, 80, 100,
  50, 40, 70, 90, 20, 110, 100, 30, 50, 90, 110, 30, 90,
  40, 80, 70])
WorkHrs = np.array([399, 121, 221, 376, 361, 224, 546,
  352, 353, 157, 160, 252, 389, 113, 435, 420, 212, 268,
  377, 421, 273, 468, 244, 342, 323])
# Fit your regression model
X = sm.add_constant(LotSize)  # Add a constant term to the
   independent variable
model = sm.OLS(WorkHrs, X).fit()

```

DRAFT

```
# Obtain residuals
residuals = model.resid

# Perform Shapiro-Wilk test for normality
statistic, p_value = shapiro(residuals)

# Print the test results
print("Shapiro-Wilk Test Statistic:", statistic)
print("p-value:", p_value)

# Interpret the results
alpha = 0.05
if p_value > alpha:
print("Residuals look Gaussian (fail to reject H0)")
else:
print("Residuals do not look Gaussian (reject H0)")
```

11.10.4 Test for Autocorrelation

Usually only meaningful for data collected in a time series or in some other sequence. In cross-sectional studies, data collection should be done so that observations are independent. A runs test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specifically designed for lack of randomness in least squares residuals, is the Durbin-Watson test.

Durbin-Watson Test

The model is:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where

$$e_i = \rho e_{i-1} + u_i, \quad |\rho| < 1$$

and

$$u_i \sim N(0, \sigma^2) \text{ independent}$$

Hypotheses:

$$H_0 : \rho = 0 \quad \text{versus} \quad H_A : \rho > 0$$

Statistic:

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Decision Rule:

$$\begin{cases} D > du & \Rightarrow \text{do not reject } H_0 \\ D < dl & \Rightarrow \text{reject } H_0 \\ dl \leq D \leq du & \Rightarrow \text{inconclusive} \end{cases}$$

Values for dl and du can be found in Table B.7 in Kutner et al.

DRAFT

Python Code: Durbin-Watson Test

To conduct the Durbin-Watson test in Python, we consider the Toluca Company example dataset given in Section 11.10.3. The Python code is provided in the following.

```
# Durbin-Watson Test
import statsmodels.api as sm
from statsmodels.stats.stattools import durbin_watson

# Define the data
LotSize = np.array([80, 30, 50, 90, 70, 60, 120, 80,
100, 50, 40, 70, 90, 20, 110, 100, 30, 50, 90, 110, 30,
90, 40, 80, 70])
WorkHrs = np.array([399, 121, 221, 376, 361, 224, 546,
352, 353, 157, 160, 252, 389, 113, 435, 420, 212, 268,
377, 421, 273, 468, 244, 342, 323])

# Fit your regression model
model = sm.OLS(WorkHrs, LotSize).fit()
print(model.summary())
# Perform the Durbin-Watson test
durbin_watson_statistic = durbin_watson(model.resid)
print("Durbin-Watson statistic:",
durbin_watson_statistic)

```

11.10.5 Tests for Non-Constancy of Variance

Tests for non-constancy of variance, also known as tests for heteroscedasticity, assess whether the variance of residuals from a regression model is constant across levels of the predictor variables. One common test is the Breusch-Pagan test. The procedure for this test is as follows:

1. Fit the regression model and obtain the residuals $\widehat{e}_i$.
2. Regress the squared residuals $\widehat{e}_i^2$ on the predictor variables x.
3. Obtain the test statistic from the resulting regression. For the Breusch-Pagan test, this statistic follows a chi-squared distribution with degrees of freedom equal to the number of predictor variables.
4. Compare the test statistic to the chi-squared distribution to determine the p-value. A small p-value indicates that the null hypothesis of constant variance is rejected, suggesting the presence of heteroscedasticity.

Breusch-Pagan Test

- Requires e_i to be independent and normally distributed.

DRAFT

- Requires large sample size.
- Can detect relationships such as

$$\log \sigma_i^2 = \gamma_0 + \gamma_1 x_i$$

Regress the squared residuals, e_i^2, against x_i and obtain SSR^* from this regression.

Hypotheses:

$$H_0 : \gamma_1 = 0 \quad \text{versus} \quad H_A : \gamma_1 > 0$$

Statistic:

$$\chi^2_{\text{BP}} = \frac{SSR^*}{2} \left(\frac{SSE}{n} \right)^2 \tag{11.15}$$

where SSR^* is the regression sum of squares when regressing $\hat{e}^2$ on x and SSE is the error sum of squares when regressing Y on x. Under H_0, the test statistic follows a χ^2-distribution.

Alternatively, the White test can be used, which is robust to various forms of heteroscedasticity and follows a similar procedure but does not assume a specific functional form of the variance.

Example: Toluca Company dataset

To conduct the Breusch-Pagan test for the Toluca Company example given in Section 11.10.3, we regress the squared residuals in Table 11.15, column 5, against x and obtain $SSR^* = 7,896,128$. We know from Figure 2.2 that $SSE = 54,825$. Hence, test statistic 11.15 is:

$$\chi^2_{\text{BP}} = \frac{7,896,128}{2} \left(\frac{54,825}{25} \right)^2 = 0.821$$

DRAFT

Table 11.15: Data on Lot Size and Work Hours

Run	(1) Lot Size (x_i)	(2) Work Hours (y_i)	(3) Mean Response ($\widehat{y}_i$)	(4) Residual (e_i)	(5) Squared Residual (e_i^2)
1	80	399	347.98	51.02	2,603.0
2	30	121	169.47	-48.47	2,349.3
3	50	221	240.88	-19.88	395.2
...	...	...	...	...	...
4	40	244	205.17	38.83	1,507.8
5	80	342	347.98	-5.98	35.8
6	70	323	312.28	10.72	114.9
Total	1,750	7,807	7,807	0	54,825

To control the α risk at 0.05, we require $\chi^2(0.95; 1) = 3.84$. Since $\chi^2_{\text{BP}} = 0.821 < 3.84$, we conclude H_0, that the error variance is constant. The p-value of this test is 0.64 so that the data are quite consistent with constancy of the error variance.

DRAFT

Python Code: Breusch-Pagan Test

To conduct the Breusch-Pagan test in Python, we consider the Toluca Company example dataset given in Section 11.10.3. The Python code is provided in the following.

```
# Breusch-Pagan Test

import numpy as np
import statsmodels.api as sm
from statsmodels.stats.diagnostic import het_breuschpagan

# Define the data
LotSize = np.array([80, 30, 50, 90, 70, 60, 120, 80, 100,
    50, 40, 70, 90, 20, 110, 100, 30, 50, 90, 110, 30, 90,
    40, 80, 70])
WorkHrs = np.array([399, 121, 221, 376, 361, 224, 546, 352,
    353, 157, 160, 252, 389, 113, 435, 420, 212, 268, 377,
    421, 273, 468, 244, 342, 323])

# Fit your regression model
X = sm.add_constant(LotSize)  # Add a constant term to the
    independent variable
```

```
model = sm.OLS(WorkHrs, X).fit()
print(model.summary())
# Perform the Breusch-Pagan test

test_results = het_breuschpagan(model.resid, model.model.
    exog)

# Extract results
bp_test_statistic = test_results[0]
bp_test_p_value = test_results[1]

print(f'Breusch-Pagan test statistic: {bp_test_statistic}')
print(f'Breusch-Pagan p-value: {bp_test_p_value}')
```

Brown-Forsythe (Modified Levene) Test

The Brown-Forsythe test, also known as the Modified Levene's test, is used to assess the equality of variances across groups. It is a robust alternative to Levene's test and is particularly useful when the data may not be normally distributed.

The Brown-Forsythe test is used to test the null hypothesis that the variances of different groups are equal. It is a modification of Levene's test, where the median is used instead of the mean to make it more robust to deviations from normality.

Procedure:

1. Arrange the residuals by increasing values of x.
2. Split the sample into two (more or less equal) groups.
   - Group 1: n_1 observations with $x \leq \tilde{x}$
   - Group 2: n_2 observations with $x > \tilde{x}$

   where $\tilde{x}$ denote the medians of x,
3. Compute $d_{i1} = |e_{i1} - \tilde{e}_1|$ and $d_{i2} = |e_{i2} - \tilde{e}_2|$, where $\tilde{e}_1$ and $\tilde{e}_2$ denote the medians of the residuals in the two groups.

Hypotheses:

$$H_0 : \sigma^2_{\text{grp1}} = \sigma^2_{\text{grp2}} \quad \text{versus} \quad H_1 : \sigma^2_{\text{grp1}} \neq \sigma^2_{\text{grp2}}$$

Statistic: Two-Sample t-Test

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where} \quad s_p^2 = \frac{\sum_{i=1}^{n_1} (d_{i1} - \bar{d}_1)^2 + \sum_{i=1}^{n_2} (d_{i2} - \bar{d}_2)^2}{n - 2}$$

DRAFT

Under H_0, the test statistic follows a t-distribution with $n-2$ degrees of freedom.

Table 11.16: Calculations for Brown Forsythe Test for Constancy of Error Variance Toluca Company Example.

Group 1					
i	**Run**	**Lot Size**	**Residual** (e_{il})	d_{il}	$(d_{il}-\bar{d}_1)^2)$
1	14	20	-20.77	0.89	1929.41
2	2	30	-48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	-60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
Total	**582.60**	—	—	**582.60**	**12,566.6**
$\tilde{e}_1 = -19.88$, $\bar{d}_1 = 44.815$					
Group 2					
i	**Run**	**Lot Size**	**Residual** (e_{i2})	d_{i2}	$(d_{i2}-\bar{d}_2)^2)$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	-34.09	31.41	8.76
12	27	120	55.21	57.89	866.10
Total	**341.40**	—	—	**341.14**	**9,610.2**
$\tilde{e}_2$ =-2.68, $\bar{d}_2$ =28.45					

We are now ready to calculate test statistic:

$$s_p^2 = \frac{12,566.6 + 9,610.2}{25-2} = 964.21$$

Hence,

$$s_p = 31.05$$

Therefore,

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{44.815 - 28.450}{31.05\sqrt{\frac{1}{13}+\frac{1}{12}}} = 1.32$$

To control the α risk at 0.05, we require $t(0.975; 23) = 2.069$. The decision rule therefore is:

DRAFT

- If $|t_{BF}| \leq 2.069$, conclude the error variance is constant.
- If $|t_{BF}|| > 2.069$, conclude the error variance is not constant.

Since $|t_{BF}| = 1.32 \leq 2.069$, we conclude that the error variance is constant and does not vary with the level of x. The two-sided p-value of this test is 0.20.

Python Code: Brown-Forsythe (Modified Levene) Test

To conduct the Brown-Forsythe (Modified Levene) Test in Python, we consider the Toluca Company example dataset given in Section 11.10.3. The Python code is provided in the following.

```
import numpy as np
import pandas as pd
import scipy.stats as stats

# Define the data
LotSize = np.array([80, 30, 50, 90, 70, 60, 120, 80, 100,
    50, 40, 70, 90, 20, 110, 100, 30, 50, 90, 110, 30, 90,
    40, 80, 70])
WorkHrs = np.array([399, 121, 221, 376, 361, 224, 546, 352,
    353, 157, 160, 252, 389, 113, 435, 420, 212, 268, 377,
    421, 273, 468, 244, 342, 323])

# Convert to DataFrame
df = pd.DataFrame({'LotSize': LotSize, 'WorkHrs': WorkHrs})

# Define bins for grouping (adjust as needed)
bins = [0, 50, 70, 90, 110, np.inf]  # Example bins
labels = ['0-50', '51-70', '71-90', '91-110', '111+']
df['Group'] = pd.cut(df['LotSize'], bins=bins, labels=labels
    )

# Compute medians for each group
group_medians = df.groupby('Group')['WorkHrs'].median()

# Compute absolute deviations from medians
df['Deviation'] = df.apply(lambda row: abs(row['WorkHrs'] -
    group_medians[row['Group']]), axis=1)

# Perform ANOVA on deviations
anova_result = stats.f_oneway(
df[df['Group'] == '0-50']['Deviation'],
df[df['Group'] == '51-70']['Deviation'],
df[df['Group'] == '71-90']['Deviation'],
df[df['Group'] == '91-110']['Deviation'],
df[df['Group'] == '111+']['Deviation']
)

```

```
print("Brown-Forsythe Test ANOVA Result:")
print(f"F-statistic: {anova_result.statistic:.4f}")
print(f"P-value: {anova_result.pvalue:.4f}")


```

11.10.6 Influential Observations, Outliers, and Cook's Distance

Cook's Distance

Cook's distance measures the influence of each observation on the fitted values of the model. It combines the effect of leverage and residual to determine the influence of each data point.

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{y}_j - \widehat{y}_{j(i)})^2}{p \cdot \text{MSE}}$$

where $\widehat{y}_j$ is the jth fitted value, $\widehat{y}_{j(i)}$ is the jth fitted value with the ith observation removed, p is the number of predictors in the model, and MSE is the mean squared error of the model.

Influential Observations

Influential observations are data points that have a large impact on the estimated coefficients of the regression model. They can significantly alter the fit of the model if removed. Influential observations are identified using Cook's distance, where observations with Cook's distance greater than $4/n$ (where n is the number of observations) are considered influential.

Outliers

Outliers are data points that deviate significantly from the rest of the data. They can affect the regression model's accuracy and should be investigated to determine if they are genuine data points or errors. Diagnostic tools such as leverage plots and Cook's distance can help identify outliers. Outliers are identified by selecting observations with Cook's distance greater than a certain threshold (here, $4/n$).

Influential Observations, Outliers, and Cook's Distance

To conduct the Influential Observations, Outliers, and Cook's Distance in Python, we consider the Toluca Company example dataset given in Section 11.10.3. The Python code is provided in the following.

DRAFT

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Define the data
LotSize = np.array([80, 30, 50, 90, 70, 60, 120, 80, 100, 50, 40, 70, 90, 20, 110, 100, 30, 50, 90, 110, 30, 90, 40, 80, 70])
WorkHrs = np.array([399, 121, 221, 376, 361, 224, 546, 352, 353, 157, 160, 252, 389, 113, 435, 420, 212, 268, 377, 421, 273, 468, 244, 342, 323])

# Create DataFrame
df = pd.DataFrame({'LotSize': LotSize, 'WorkHrs': WorkHrs})

# Add a constant column for intercept
X = sm.add_constant(df['LotSize'])
y = df['WorkHrs']

# Fit the linear regression model
model = sm.OLS(y, X).fit()

# Get influence measures
influence = model.get_influence()
cooks_d = influence.cooks_distance[0]

# Create a DataFrame for Cook's Distance
cooks_df = pd.DataFrame({
  'LotSize': df['LotSize'],
  'WorkHrs': df['WorkHrs'],
  'Cook\'s Distance': cooks_d
})

# Print the Cook's Distance values
print("Cook's Distance:")
print(cooks_df)

# Plot Cook's Distance
plt.figure(figsize=(10, 6))
plt.stem(np.arange(len(cooks_d)), cooks_d, basefmt=" ", use_line_collection=True)
plt.xlabel('Observation Index')
plt.ylabel("Cook's Distance")
plt.title("Cook's Distance for Each Observation")
plt.show()

# Identify influential observations (threshold typically 4/n)
```

```
threshold = 4 / len(cooks_d)
influential = cooks_df[cooks_df['Cook\'s Distance'] >
    threshold]

print("Influential Observations:")
print(influential)
```

11.10.7 Multicollinearity

Multicollinearity is a common issue in regression analysis. It occurs when predictor variables in a regression model are highly correlated. Multicollinearity can lead to inaccurate estimates of the regression coefficients and their standard errors.

> **Multicollinearity:** Multicollinearity is a statistical phenomenon where two or more independent variables in a regression model are highly correlated, leading to unreliable estimates of regression coefficients.

Causes of Multicollinearity

Multicollinearity can be caused by:

- **High Correlation Among Predictors:** When independent variables are highly correlated with each other.
- **Redundant Variables:** Including variables that are redundant or linear combinations of other predictors.
- **Inclusion of Polynomial or Interaction Terms:** When polynomial terms or interaction terms are added without centering the variables.
- **Data Collection Issues:** Poorly designed experiments or data collection methods that capture similar information across multiple variables.

What is the problem if one feature is a linear combination of other features? If one feature is a linear combination of other features, it indicates multicollinearity. Let $\mathbf{X}$ be the design matrix of dimension $n \times p$, where n is the number of observations and p is the number of predictors. Each row of $\mathbf{X}$ corresponds to an observation, and each column corresponds to a predictor variable.

The multiple linear regression model can be represented as:

$$\vec{\boldsymbol{Y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{e}}$$

where:

DRAFT

- $\vec{\boldsymbol{Y}}$ is the $n \times 1$ vector of response variable values,
- $\vec{\boldsymbol{\beta}}$ is the $p \times 1$ vector of regression coefficients,
- $\vec{\boldsymbol{e}}$ is the $n \times 1$ vector of errors.

The estimator of $\vec{\boldsymbol{\beta}}$ is

$$\vec{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{\boldsymbol{Y}}$$

If one feature is a linear combination of other features, then the rank of the design matrix $\mathbf{X}$ is less then the number of predictors. That is,

$$\text{rank}(\mathbf{X}) < p.$$

If the rank of $\mathbf{X}$ is less than the number of predictors p, then the determinant of $\mathbf{X}^T\mathbf{X}$ is close to zero. If one feature is an exact linear combination of other features, then the determinant of $\mathbf{X}^T\mathbf{X}$ is exactly zero. Consequently, $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exist, which implies that the Ordinary Least Squares (OLS) estimator $\vec{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{\boldsymbol{Y}}$ and $\text{Var}(\vec{\hat{\boldsymbol{\beta}}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$ are undefined or cannot be computed in such cases.

In scenarios where multicollinearity is present, leading to a nearly singular $\mathbf{X}^T\mathbf{X}$, the OLS estimator $\vec{\hat{\boldsymbol{\beta}}}$ can become very large and provide unreliable coefficient estimates. Additionally, the variance of $\vec{\hat{\boldsymbol{\beta}}}$ becomes very large. This large variance results in small t-ratios, which undermines the validity of hypothesis tests. Furthermore, the large variance leads to wide confidence intervals for the true parameters, affecting the accuracy of inference about the parameters.

Therefore, when one feature is a linear combination of other features, it introduces significant problems for statistical inference and makes it challenging to interpret the effects of individual predictors on the response variable.

Practical Consequences of Multicollinearity

Some practical consequences of multicollinearity include:

1. **Large Variances and Covariances**: Although Ordinary Least Squares (OLS) estimators are Best Linear Unbiased Estimators (BLUE), multicollinearity leads to large variances and covariances of the estimators. This makes precise estimation of coefficients difficult.

2. **Wider Confidence Intervals**: Due to the large variances of the estimators, the confidence intervals for the coefficients tend to be much wider. This often leads to the acceptance of the null hypothesis that the true population coefficient is zero, even when it may not be.

DRAFT

3. **Insignificant t-Ratios**: The large variances result in t-ratios that tend to be statistically insignificant. This makes it challenging to determine whether individual coefficients are significantly different from zero.

4. **High R^2 with Insignificant Coefficients**: Despite the t-ratios of one or more coefficients being statistically insignificant, the overall measure of goodness of fit, R^2, can still be very high. This can create a misleading impression of model fit.

5. **Sensitivity to Data Changes**: OLS estimators and their standard errors can be highly sensitive to small changes in the data. This instability further complicates the interpretation and reliability of the regression results.

Variance Inflation Factor (VIF)

The **Variance Inflation Factor (VIF)** is a measure used to quantify how much the variance of an estimated regression coefficient increases due to multicollinearity. For a given predictor X_j in a multiple regression model, the VIF for each predictor variable, denoted as VIF_j for jth predictor, is calculated as:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination from regressing the jth predictor variable on all the other predictor variables.

To check for multicollinearity in the model:

1. Calculate the Variance Inflation Factor (VIF) for each predictor variable, $x_{1i}, x_{2i}, \ldots, x_{pi}$.

2. A commonly used rule of thumb is that **if the VIF for any predictor variable exceeds 10**, multicollinearity may be a concern. If there is no collinearity between x_2 and x_3 for three variable regression model, VIF will be 1.

3. Repeat steps 1-2 for each predictor variable in the model.

Python Code: Variance Inflation Factor (VIF)

To calculate VIF in Python using the `statsmodels` library, we can use the following code for hypothetical data:

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import
    variance_inflation_factor

```

DRAFT

```
# Example DataFrame
df = pd.DataFrame({
  'X1': [1, 2, 3, 4, 5],
  'X2': [2, 4, 6, 8, 10],
  'X3': [5, 4, 3, 2, 1]
})

# Add a constant column for intercept
X = sm.add_constant(df)

# Calculate VIF for each feature
vif = pd.DataFrame()
vif['Feature'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i
    in range(X.shape[1])]

print(vif)


```

Tolerance of Limit (TOL)

The **Tolerance of Limit (TOL)** is a measure used to quantify how much of the variance of a predictor is not explained by the other predictors in a regression model. It is defined as:

$$\text{TOL}_j = \frac{1}{\text{VIF}_j} = 1 - R_j^2$$

When $R_j^2 = 1$ (i.e., perfect collinearity), $\text{TOL}_j = 0$ and when $R_j^2 = 0$ (i.e., no collinearity whatsoever), TOL_j is 1. Because of the intimate connection between VIF and TOL, one can use them interchangeably

Python Code: Tolerance of Limit (TOL)

To calculate tolerance in Python, we can use the relationship with VIF, as VIF is the reciprocal of tolerance. Here's how we can calculate it using the same example from before:

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import
    variance_inflation_factor

# Example DataFrame
df = pd.DataFrame({
  'X1': [1, 2, 3, 4, 5],
  'X2': [2, 4, 6, 8, 10],
  'X3': [5, 4, 3, 2, 1]
```

DRAFT

```
})

# Add a constant column for intercept
X = sm.add_constant(df)

# Calculate VIF for each feature
vif = pd.DataFrame()
vif['Feature'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i
    in range(X.shape[1])]
vif['Tolerance'] = 1 / vif['VIF']

print(vif)
```

Remedial Measures for Multicollinearity

To address multicollinearity, several remedial measures can be employed:

1. **Remove Highly Correlated Predictors**:
   - Identify and remove one or more predictors that are highly correlated with others.
   - Use correlation matrices, Variance Inflation Factor (VIF), or other diagnostics to detect high multicollinearity.
2. **Combine Predictors**:
   - Combine correlated predictors into a single composite variable.
   - Use techniques such as Principal Component Analysis (PCA) to create a new set of uncorrelated variables.
3. **Apply Regularization Techniques**:
   - Regularization methods can help reduce multicollinearity by adding a penalty to the size of the coefficients.
   - Techniques include:
     - **Ridge Regression (L_2 Regularization)**: Adds a penalty proportional to the square of the magnitude of coefficients.
     - **Lasso Regression (L_1 Regularization)**: Adds a penalty proportional to the absolute value of coefficients, which can also perform feature selection by shrinking some coefficients to zero.
4. **Use Principal Component Analysis (PCA)**:
   - PCA transforms the predictors into a set of orthogonal components.

DRAFT

- Replace the original predictors with the principal components to address multicollinearity.

5. **Increase Sample Size**:
   - A larger sample size can reduce the variance of the coefficient estimates and mitigate the effects of multicollinearity.
   - Collect more data if feasible to improve the stability of the regression model.
6. **Center the Variables**:
   - Centering involves subtracting the mean of each predictor from the predictor values.
   - This is particularly useful when dealing with polynomial or interaction terms.
7. **Drop or Transform Variables**:
   - Drop variables that do not contribute significantly to the model or transform variables to reduce multicollinearity.
   - Conduct feature selection or use transformations to address issues.

11.10.8 Exercises

1. Define multicollinearity. How can it affect the results of a multiple linear regression model, and how can it be detected?
2. Explain how you would use a residual plot to assess the fit of a multiple linear regression model. What patterns in the residual plot might suggest problems with the model?
3. List the key assumptions of linear regression. For each assumption, provide a brief explanation of why it is important for the validity of the regression model.
4. Given a dataset with a linear regression model fitted, describe how you would check each of the assumptions (linearity, independence, homoscedasticity, normality of errors).
5. Explain the Durbin-Watson test and how it is used to test for autocorrelation.
6. Explain how to perform the Breusch-Pagan test and the White test for heteroscedasticity. Interpret the results of these tests and discuss how they affect the validity of the regression model.

DRAFT

7. How would you handle outliers and influential observations? Discuss methods such as robust regression or data transformation.

8. Explain how to detect multicollinearity in a regression model using Variance Inflation Factor (VIF) or condition indices.

9. If multicollinearity is present, what strategies can you use to address it? Discuss methods like variable selection or principal component analysis.

10. How would you address an influential observation identified through Cook's Distance?

11.11 Concluding Remarks

This chapter has equipped you with essential techniques in correlation and regression analysis, crucial for the practice of data science. By examining scatter diagrams, covariance, and correlation coefficients, we laid the groundwork for understanding the relationships between variables in a dataset.

The exploration of regression analysis, including both simple and multiple linear regression models, has highlighted key aspects such as model assumptions, coefficient interpretation, and model fit evaluation. These tools are fundamental for building predictive models, making data-driven decisions, and deriving actionable insights from complex datasets.

The inclusion of Python code examples throughout the chapter bridges the gap between theoretical concepts and practical implementation, demonstrating how to apply these techniques in real-world data science scenarios. Mastering these methods will enhance your ability to analyze data, validate findings, and contribute to evidence-based decision-making across various domains.

In summary, a solid grasp of correlation and regression analysis not only improves our analytical skills but also enables us to leverage data effectively to drive strategic decisions and solve complex problems in the field of data science.

11.12 Chapter Exercises

1. What is meant by correlation and regression? Distinguish between simple linear regression and multiple linear regression.

2. Explain the difference between linear and nonlinear regression models.

3. Suppose data were collected from a sample of 12 semesters from a university bookstore. For the ith semester in the sample, x_i is the number of introductory biology courses offered and y_i is the total sales of introductory biology textbooks (in thousands of dollars).

DRAFT

Table 11.17: Introductory Biology Courses and Textbook Sales Data for Different Semesters

Semesters i	**No. of Courses** x_i	**Textbook Sales (\$1000s)** y_i
1	4	12
2	5	15
3	5	17
4	6	19
5	6	21
6	7	22
7	7	23
8	8	24
9	9	26
10	9	28
11	10	29
12	10	31

(i). Show the relationship between the number of introductory biology courses and the total sales of introductory biology textbooks. Make a comment on the diagram.

(ii). Write down the regression model for this example, and mention the assumptions of the model.

(iii). Find the least square estimates and write the estimated regression model. Interpret the results.

(iv). Draw the regression line on the scatter diagram.

(v). Predict total textbook sales for a semester in which 8 introductory biology courses are offered.

(vi). Find the value of the standard error of the estimates.

4. A study was conducted to investigate the effect of a new drug on blood pressure. The dataset contains information on the age of patients, their baseline blood pressure, and their blood pressure after 6 weeks of treatment with the drug. The following data has been collected:

DRAFT

Patient (years)	**Age (mm Hg)**	**Baseline Blood Pressure (mm Hg)**	**Blood Pressure after 6 weeks** (y)
1	45	140	130
2	50	150	135
3	60	160	140
4	55	155	145
5	65	170	150
6	70	175	155
7	75	180	160
8	80	185	165
9	85	190	170
10	90	195	175

(i) Formulate a linear regression model to study the relationship between age (x_1) and baseline blood pressure (x_2) as independent variables and the blood pressure after 6 weeks (y) as the dependent variable.

(ii) State the assumptions of the linear regression model in the context of this study.

(iii) Estimate the regression coefficients using the least squares method. Interpret the coefficients for age and baseline blood pressure.

(iv) Using the model obtained in part (iii), predict the blood pressure after 6 weeks for a 65-year-old patient with a baseline blood pressure of 165 mm Hg.

(v) Discuss the potential impact of multicollinearity between age and baseline blood pressure on the model's estimates.

(vi) Calculate the standard error of the estimates and discuss the precision of the regression coefficients.

(vii) Perform a hypothesis test to determine whether age is a significant predictor of blood pressure after 6 weeks. Use a 5% significance level.

(viii) Construct a 95% confidence interval for the coefficient of baseline blood pressure. Interpret the interval.

(ix) Create a residual plot and comment on the model's assumptions regarding homoscedasticity and normality of errors.

(x) Discuss the limitations of using this linear regression model for predicting blood pressure after 6 weeks and suggest possible improvements.

DRAFT

5. An engineer is studying the tensile strength of a new composite material. The dataset contains information on the fiber content of the composite, the curing temperature, and the measured tensile strength. The following data has been collected:

Sample (%)	**Fiber Content (°C)**	**Curing Temperature (MPa)**	**Tensile Strength** (y)
1	30	150	450
2	35	160	470
3	40	170	480
4	45	180	490
5	50	190	510
6	55	200	530
7	60	210	540
8	65	220	550
9	70	230	560
10	75	240	570

(i) Develop a multiple linear regression model where tensile strength (y) is the dependent variable, and fiber content (x_1) and curing temperature (x_2) are independent variables.

(ii) Explain the assumptions of the multiple linear regression model in the context of this engineering problem.

(iii) Estimate the regression coefficients using the least squares method. Interpret the coefficients for fiber content and curing temperature.

(iv) Using the model obtained in part (iii), estimate the tensile strength of a composite with 50% fiber content cured at 200°C.

(v) Calculate the coefficient of determination (R^2) and interpret its value in terms of the model's goodness-of-fit.

(vi) Conduct a hypothesis test to assess whether fiber content has a statistically significant effect on tensile strength. Use a 5% significance level.

(vii) Construct a 95% confidence interval for the coefficient of curing temperature. What does this interval indicate about the relationship between curing temperature and tensile strength?

(viii) Plot the predicted tensile strength against the actual tensile strength and comment on the model's predictive accuracy.

DRAFT

(ix) Discuss how changes in fiber content and curing temperature might interact to affect tensile strength. Could an interaction term be included in the model?

(x) Analyze the residuals from the model to check for any violations of the regression assumptions, such as non-linearity or heteroscedasticity.

6. Suppose age (days), birthweight (oz), and SBP are measured for 16 infants and the data are as shown in Table 11.18. What is the relationship between infant systolic blood pressure (SBP) and their age and birthweight? Can we predict SBP based on these factors?

Table 11.18: Sample data for infant blood pressure, age, and birthweight for 16 infants

i	Age (days) (x_1)	Birthweight (oz) (x_2)	SBP (mm Hg) (y)
1	3	135	89
2	4	120	90
3	3	100	83
4	2	105	77
5	4	130	92
6	5	125	98
7	2	125	82
8	3	105	85
9	5	120	96
10	4	90	95
11	2	120	80
12	3	95	79
13	3	120	86
14	4	150	97
15	3	160	92
16	3	125	88

Based on the data provided in Example 6, answer the following questions:

(a) **Descriptive Statistics:**

i. Calculate the mean and standard deviation for the variables:

DRAFT

- Age (days) (x_1)
- Birthweight (oz) (x_2)
- SBP (mm Hg) (y)

ii. What do the mean and standard deviation tell you about the distribution of each variable?

(b) **Scatter Plots:**

i. Create a scatter plot of Age (days) (x_1) vs. SBP (mm Hg) (y).

ii. Create a scatter plot of Birthweight (oz) (x_2) vs. SBP (mm Hg) (y).

iii. Visually assess the relationship between SBP and each predictor. Is the relationship linear or nonlinear?

(c) **Correlation Analysis:**

i. Calculate the Pearson correlation coefficient between Age (x_1) and SBP (y).

ii. Calculate the Pearson correlation coefficient between Birthweight (x_2) and SBP (y).

iii. Interpret the correlation coefficients. Which variable is more strongly correlated with SBP?

(d) **Simple Linear Regression:**

i. Perform a simple linear regression to predict SBP (y) based on Age (x_1). Write down the regression equation.

ii. Perform a simple linear regression to predict SBP (y) based on Birthweight (x_2). Write down the regression equation.

iii. Interpret the slope of each regression line. What does the slope tell you about the relationship between each predictor and SBP?

(e) **Multiple Linear Regression:**

i. Perform a multiple linear regression using Age (x_1) and Birthweight (x_2) as predictors of SBP (y).

ii. Write down the multiple regression equation.

iii. Interpret the coefficients of Age and Birthweight in the context of predicting SBP.

(f) **Model Comparison:**

i. Calculate the R^2 value for the simple linear regression models from Question 6d.

ii. Calculate the R^2 value for the multiple linear regression model from Question 6e.

iii. Compare the R^2 values. Which model explains more variance in SBP?

(g) **Prediction:**

DRAFT

i. Using the multiple regression model from Question 6e, predict the SBP for an infant who is 4 days old with a birthweight of 115 oz.

ii. Compare this prediction to the SBP predicted by each of the simple linear regression models (age-based and birthweight-based).

(h) **Model Assumptions:**

i. Discuss the assumptions underlying the multiple linear regression model (e.g., linearity, independence, homoscedasticity, normality).

ii. How would you check these assumptions with the given data?

iii. Check for multicollinearity between Age and Birthweight. What are the potential implications of multicollinearity on your regression model?

Summarize the relationship between infant systolic blood pressure (SBP) and the predictors (Age and Birthweight). Based on your analysis, do you think Age and Birthweight are sufficient to predict SBP? Suggest other factors that might be important to include in the model.

7. Consider a dataset containing information about the annual revenue of a tech company, the number of employees, and the research and development (R&D) expenditure. The following data has been collected:

Company (1000s)	**Number of Employees (Millions of Dollars)**	**R&D Expenditure (Millions of Dollars)**	**Annual Revenue** (y)
1	5	8	120
2	12	15	180
3	9	10	150
4	14	20	200
5	18	25	250
6	22	28	280
7	25	30	320
8	30	35	350
9	28	32	340
10	35	40	400

(i) Using the above data, formulate a multiple linear regression model where the annual revenue (y) is the dependent variable, and both the number of employees (x_1) and the R&D expenditure (x_2) are independent variables.

(ii) Write down the assumptions of the multiple linear regression model.

(iii) Estimate the regression coefficients for the model formulated in Question 1 using the least squares method. Interpret the coefficients for both the number of employees and R&D expenditure.

DRAFT

(iv) Based on the regression model obtained in part (iii), what would be the estimated annual revenue for a company with 20,000 employees and an R&D expenditure of $30 million?

(v) Calculate the standard error of the estimates for the model obtained in Question (iii).

(vi) How would you evaluate the goodness-of-fit for the regression model? What statistical metrics would you consider, and why?

(vii) Create scatter plots to show the relationship between:

- Number of employees and annual revenue.
- R&D expenditure and annual revenue.

Overlay the regression lines on these plots and comment on the observed relationships.

(viii) If a new company is planning to hire 25,000 employees and spend $35 million on R&D, use the regression model from Question (iii) to predict the annual revenue.

(ix) Discuss the limitations of using the linear regression model in this context. What are some factors not considered by the model that could affect the accuracy of your predictions?

(x) Test the significance of each regression coefficient at the 5% significance level. Clearly state the null and alternative hypotheses, the test statistic, and your conclusion.

(xi) Construct a 95% confidence interval for the coefficients of the number of employees and R&D expenditure. Interpret the intervals.

8. **Grocery Retailer:** A large, national grocery retailer tracks productivity and costs of its facilities closely. Data given in Table 11.19 were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are:

- The number of cases shipped (X_1)
- The indirect costs of the total labor hours as a percentage (X_2)
- A qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3)
- The total labor hours (Y)

(i). Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

(ii). Write a multiple regression model to the data for three predictor variables. State the estimate regression function.

DRAFT

(iii). Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

(iv). Plot the residuals against Y, X_1, X_2, X_3, and X_1X_2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

(v). Prepare a time plot of the residuals. Is there any indication that the error terms are correlated? Discuss.

(vi). Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = 0.01$. State the decision rule and conclusion.

(vii). Test whether there is a regression relation, using a level of significance of 0.05. State the alternatives, decision rule, and conclusion. What does your test result imply about β_1, β_2, and β_3? What is the P-value of the test?

(viii). Calculate the coefficient of multiple determination R^2. How is this measure interpreted here?

(ix). From separate shipments with the following characteristics must be processed next month:

X_1	X_2	X_3
230,000	7.50	0
250,000	7.30	0
280,000	7.10	0
340,000	6.90	0

Management desires predictions of the handling times for these shipments so that the actual handling times can be compared with the predicted times to determine whether any are out of line. Develop the needed predictions, using the most efficient approach and a family confidence coefficient of 95%.

(x). Three new shipments are to be received, each with $X_1 = 282,000$, $X_2 = 7.10$, and $X_3 = 0$.

(a). Obtain a 95% prediction interval for the mean handling time for these shipments.

(b). Convert the interval obtained in part (a) into a 95% prediction interval for the total labor hours for the three shipments.

DRAFT

Table 11.19: Grocery Retailer

y	x1	x2	x3
4264	305657	7.17	0
4496	328476	6.2	0
4317	317164	4.61	0
4292	366745	7.02	0
4945	265518	8.61	1
4325	301995	6.88	0
4110	269334	7.23	0
4111	267631	6.27	0
4161	296350	6.49	0
4560	277223	6.37	0
4401	269189	7.05	0
4251	277133	6.34	0
4222	282892	6.94	0
4063	306639	8.56	0
4343	328405	6.71	0
4833	321773	5.82	1
4453	272319	6.82	0
4195	293880	8.38	0
4394	300867	7.72	0
4099	296872	7.67	0
4816	245674	7.72	1
4867	211944	6.45	1
4114	227996	7.22	0
4314	248328	8.5	0
4289	249894	8.08	0
4269	302660	7.26	0
4347	273848	7.39	0
4178	245743	8.12	0

Continued on next page

DRAFT

Continued from previous page

y	x1	x2	x3
4333	267673	6.75	0
4226	256506	7.79	0
4121	271854	7.89	0
3998	293225	9.01	0
4475	269121	8.01	0
4545	322812	7.21	0
4016	252225	7.85	0
4207	261365	6.14	0
4148	287645	6.76	0
4562	289666	7.92	0
4146	270051	8.19	0
4555	265239	7.55	0
4365	352466	6.94	0
4471	426908	7.25	0
5045	369989	9.65	1
4469	472476	8.2	0
4408	414102	8.02	0
4219	302507	6.72	0
4211	382686	7.23	0
4993	442782	7.61	1
4309	322303	7.39	0
4499	290455	7.99	0
4186	411750	7.83	0
4342	292087	7.77	0

DRAFT

Chapter 12

Applied Time Series Analysis and Forecasting

DRAFT

12.1 Introduction

Time series analysis is a cornerstone of data science, essential for understanding and forecasting temporal dynamics in data. Unlike static datasets, time series data are collected over time, providing a sequence of observations that reflect changes and patterns occurring at different intervals. This temporal structure introduces unique challenges and opportunities for analysis, making time series a critical area of study for researchers and practitioners across various domains.

In essence, a time series is a series of data points indexed in time order, where each data point represents an observation at a specific time. These observations can span seconds, minutes, hours, days, months, or even years, depending on the application. The primary goal of time series analysis is to uncover underlying patterns such as trends, seasonality, and cyclic behavior, and to use these insights for forecasting future values.

The process of analyzing time series data begins with visualizing the data to identify patterns and anomalies. For instance, the **AirPassengers** dataset, which records the monthly totals of international airline passengers from 1949 to 1960, serves as a classic example for illustrating key concepts in time series analysis. Through this dataset, we can observe seasonality and trends that are typical in time series data.

Understanding the importance of time series analysis is crucial. It enables organizations and researchers to make informed decisions based on historical data trends and patterns, optimize processes, and predict future events with greater accuracy. Time series forecasting, a primary application of time series

analysis, allows for the development of models that can anticipate future values based on past observations.

In this chapter, we will explore the foundational concepts and techniques used in time series analysis. We will differentiate between time series and other types of data such as cross-sectional and pooled data, and we will examine various methods and models used to analyze and forecast time series data. By delving into these topics, we aim to provide a comprehensive understanding of how to effectively handle and interpret time series data in practical applications.

12.2 What is a Time Series?

Suppose you manage a coffee shop and aim to optimize inventory levels by forecasting monthly sales. Relying on data from just one month provides insufficient information for accurate forecasting. However, by leveraging data collected over several years, data scientists can apply time series analysis techniques to discern trends, seasonal patterns, and other key features, thereby improving forecast accuracy and strategic planning.

Time Series Analysis: A set of statistical techniques for analyzing data collected at regular intervals (e.g., daily, monthly, annually) to identify trends, patterns, and seasonal effects, and to make forecasts.

A time series is a sequence of observations on a variable measured at successive points in time or over successive periods of time. This sequence is often organized in a time-ordered fashion, meaning that each data point is indexed according to its time of measurement. Most commonly, a time series consists of data collected at **equally spaced intervals**, such as hourly, daily, weekly, monthly, or annually.

Given this structure, time series analysis becomes a powerful tool for extracting meaningful insights from temporal data. By analyzing the sequence of data points, we can forecast future events and make informed decisions based on historical patterns. The process involves developing models that capture the underlying patterns and trends in the data, aiming to understand the causes behind observed behaviors.

Ultimately, the primary objective of time series analysis is to create mathematical models that offer plausible descriptions of the data, enabling accurate predictions and a deeper understanding of the temporal dynamics at play. Here's how time series analysis is applied in different domains:

DRAFT

Finance

In finance, time series analysis is extensively used for forecasting stock prices, analyzing market trends, and managing risk. Portfolio managers rely on time series analysis to understand historical performance and forecast returns, optimizing asset allocation to achieve desired financial goals.

Healthcare

In healthcare, time series analysis aids in monitoring and predicting patient health metrics over time. Medical researchers use it to analyze trends in disease incidence, track the effectiveness of treatments, and predict future healthcare needs. For instance, time series models can forecast the progression of chronic diseases, anticipate patient admissions, and optimize resource allocation in hospitals. Wearable health devices also generate time series data that can be analyzed to monitor vital signs and detect anomalies.

Retail

Retailers use time series analysis to forecast demand, optimize inventory, and plan marketing campaigns. Retail analysts analyze historical sales data to predict future demand for products, ensuring that inventory levels are aligned with expected sales. Time series models help identify seasonal trends, such as increased sales during holiday seasons, allowing for targeted promotions and stock adjustments. Supply chain managers use these insights to streamline procurement and reduce excess inventory costs.

Energy

In the energy sector, time series analysis is crucial for managing supply and demand, forecasting energy consumption, and optimizing resource allocation. **Utility companies** use it to predict electricity demand based on historical consumption patterns, weather conditions, and other factors. **Energy traders** rely on time series forecasts to make informed decisions about energy contracts and pricing. Time series analysis also helps in **monitoring equipment performance** and predicting maintenance needs to avoid downtime.

Transportation and Logistics

Time series analysis helps optimize routes, manage fleet operations, and forecast travel demand in transportation and logistics. **Logistics companies** use it to predict shipping volumes and plan warehouse operations accordingly. **Public transportation authorities** analyze historical ridership data to improve scheduling and service levels. **Traffic management systems** use time series data to forecast congestion patterns and implement traffic control measures effectively.

DRAFT

Climate Science

In climate science, time series analysis is used to study long-term climate trends, model weather patterns, and predict natural disasters. **Climate scientists** analyze historical weather data to understand changes in temperature, precipitation, and other climate variables. Time series models help forecast extreme weather events, such as hurricanes or heatwaves, and assess their potential impact on various regions. **Environmental policymakers** use these insights to develop strategies for climate adaptation and mitigation.

Economics

Economists use time series analysis to study economic indicators, forecast economic growth, and analyze market cycles. **Macroeconomic analysts** examine historical data on GDP, unemployment rates, and inflation to model economic trends and predict future economic conditions. Time series analysis also helps in assessing the impact of economic policies and external shocks on national and global economies.

In each of these fields, time series analysis enables professionals to make data-driven decisions by identifying trends, forecasting future values, and understanding complex temporal patterns. This application of time series techniques enhances strategic planning, operational efficiency, and decision-making across diverse industries.

DRAFT

12.2.1 Time series Plot

A time series plot is a graphical representation of data points in a time sequence, where the x-axis represents time and the y-axis represents the observed values. This plot helps to identify whether the data is trending upwards, downwards, or remaining constant. This plot is used to identify Periodic patterns (e.g., yearly cycles). Time series plots are often the first step in exploring time-dependent data before moving on to more complex analyses, like decomposition or forecasting models.

12.2.2 Example: AirPassengers dataset

The **AirPassengers** dataset is a classic time series dataset that records the **monthly** totals of international airline passengers (in thousands) from January 1949 to December 1960. It is often used in time series analysis and forecasting examples due to its clear trends and seasonality patterns. Here is an example of how to create a time series plot for the **AirPassengers** dataset in Python:

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the AirPassengers dataset
```

```
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/
    master/airline-passengers.csv'
data = pd.read_csv(url, index_col='Month', parse_dates=True)

# Plot the time series
plt.figure(figsize=(10, 6))
plt.plot(data, color='blue', marker='o', linestyle='-')
plt.title("Monthly Totals of International Airline
    Passengers (1949 - 1960)")
plt.xlabel("Year")
plt.ylabel("Number of Passengers (in thousands)")
plt.grid(True)
plt.show()
```

Listing 12.1: Time Series Plot of Monthly Totals of International Airline Passengers.

The time series plot includes markers for individual data points and a line to connect them. For the example, the x-axis represents time (years), and the y-axis represents the number of airline passengers. The time series plot is presented in Figure 12.1. The plot shows a clear upward trend, indicating that the number of airline passengers increased over time. Repeating patterns can be observed, with certain months showing consistently higher passenger numbers, highlighting seasonal effects in the data.

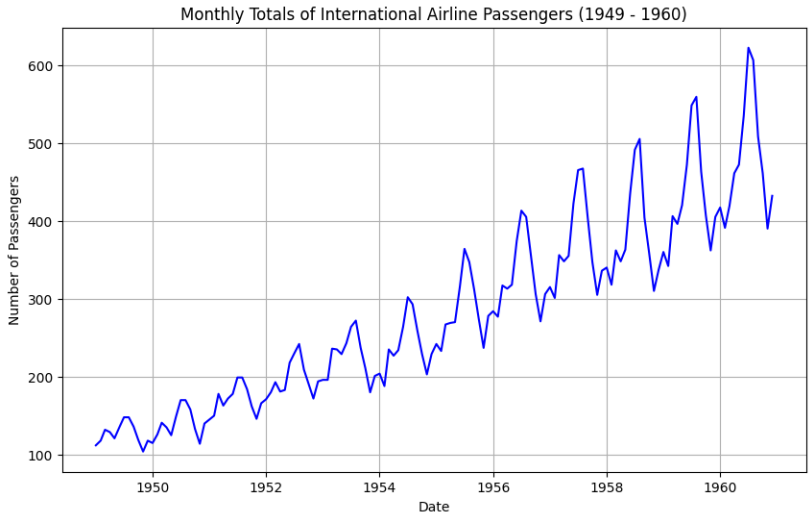


Figure 12.1: Time Series Plot for the AirPassengers Dataset

12.2.3 Importance of Time Series Analysis

Time series analysis is critical in various fields, including economics, finance, meteorology, engineering, and many others. It allows analysts to:

- Identify patterns such as trends and seasonality.
- Model the underlying process generating the data.
- Forecast future values based on historical data.
- Detect outliers or unusual patterns.

12.2.4 Time Series Data vs. Cross-Sectional Data

Time series data provides insights into how a variable evolves over time, offering a dataset that reveals the variations and trends of the variable being studied. For example, time series data might include monthly returns over the past five years or salary data collected month-by-month across a city.

In contrast, cross-sectional data is obtained by capturing observations from multiple individuals or entities at the same point in time. This type of data provides a snapshot of various subjects simultaneously. Examples include the year-end book values per share for all New York Stock Exchange-listed companies in 2003, or a dataset of salaries across a city collected at a single moment.

Time Series Data: Data collected at successive time points for a single entity (e.g., monthly stock prices for a company).

Cross-Sectional Data: Data collected at a single point in time across multiple entities (e.g., survey results from different households at one moment).

12.2.5 Pooled Data vs. Longitudinal Data

Pooled data, also known as panel data, combines observations from multiple time points with those collected at a single point in time. For instance, a salary dataset might include data gathered both at one specific moment and over several different periods.

In contrast, longitudinal data consists of observations on various characteristics of a single observational unit collected at different times. For example, this might involve tracking a company's financial ratios over a span of 10 years.

Pooled Data: Data that combines cross-sectional data from multiple time periods or multiple entities into a single dataset.

DRAFT

Longitudinal Data: Data collected from the same entities over multiple time periods, allowing for the study of changes over time.

12.2.6 Types of Time Series

There are two main types of time series: **discrete** and **continuous**. A time series is considered discrete when observations are taken only at specific times, typically at equally spaced intervals. In contrast, a time series is categorized as continuous when observations are recorded continuously over a time interval, such as a binary process monitored throughout time.

Discrete time series can arise from continuous time series in a couple of ways. One method is sampling, where values from a continuous time series are recorded at equal intervals to produce a discrete time series. It is crucial to choose the sampling interval carefully to retain as much information as possible. Another method is aggregation, where values from a continuous time series are averaged over specified intervals, such as monthly exports or daily rainfall measurements. In this chapter, we will focus more on discrete time series.

Discrete Time Series Analysis: The study of data points collected at specific, evenly spaced intervals over time.

Continuous Time Series Analysis: The examination of data recorded continuously throughout a time interval, providing a smooth and uninterrupted view of temporal changes.

Continuous time series involve observations that are recorded at every possible moment within a time interval, providing a smooth and uninterrupted data flow. This type of time series is particularly useful for capturing phenomena that evolve in a continuous manner, such as the real-time monitoring of heart rate during an exercise session. In this example, the heart rate is measured continuously, allowing for detailed insights into how it fluctuates with every moment of physical activity. Although in practice, such data might be sampled at high frequencies and then smoothed to approximate continuous behavior, it still provides a rich representation of temporal changes. Understanding continuous time series is crucial for applications that require fine-grained analysis and modeling, where capturing the nuances of continuous changes is essential for accurate prediction and control.

12.2.7 Time Series Forecasting

Time series **forecasting**, a modern approach in data analysis, involves making predictions about future values in a sequence of time-ordered data points based on historical data. This process is referred to as **extrapolation** in classical statistics, where the goal is to extend the data beyond its current range.

DRAFT

Forecasting specifically focuses on using models fit to past observations to estimate future outcomes, a process that inherently assumes the future is entirely unknown and must be inferred from past events.

Time series forecasting: The process of predicting future values in a time-ordered sequence based on historical data.

12.2.8 Preliminary Adjustments Before Analyzing Time Series

Before you start analyzing a time series, it is important to make some adjustments to the raw data. These adjustments help ensure that your analysis reflects true trends and patterns. Here are the main adjustments to consider:

1. **Adjustment for Calendar Variation**
    - Calendar variation happens because different months and seasons have different numbers of working days, weekends, and holidays. For example, sales might spike during holiday seasons, making it hard to see the actual trends.
    - **How to Adjust**:
        - *Seasonal Decomposition*: Break down the data to separate seasonal patterns from actual trends.
        - *Working Day Adjustment*: Normalize data based on the number of working days in a month.
    - **Example**: If retail sales in December show a spike due to Christmas, analyzing only December without adjusting for this holiday effect could misrepresent sales trends for the entire year.
2. **Adjustment for Population Changes**
    - Population changes can greatly impact data. For instance, a growing population might naturally lead to higher sales or more healthcare needs.
    - **How to Adjust**:
        - *Per Capita Measurement*: Express values per person (e.g., per 1,000 people) to allow for comparisons across different population sizes.
        - *Weighting*: Adjust data based on population estimates to get a clearer picture.
    - **Example**: If a city's population grows from 100,000 to 120,000 and the number of hospital visits increases from 10,000 to 12,000, the visits per 1,000 people remain at 100, indicating that there

DRAFT

has been no increase in hospital usage per person despite the increase in total visits and population.

3. **Adjustment for Price Changes**

   - Prices can change over time due to inflation or deflation, affecting the value of monetary data. This can lead to misleading conclusions if not addressed.
   - **How to Adjust**:
     - *Use Price Indices*: Apply a price index like the Consumer Price Index (CPI) to convert nominal values to real values, reflecting true purchasing power.
     - *Inflation Adjustment*: Adjust historical prices to current values to make comparisons more accurate.
   - **Example**: If a product's price increased from \$100 to \$120 over a year, but inflation was 10%, the real price increase is actually only \$110 in terms of purchasing power, not \$120. Therefore, even though the nominal price increased by \$20 (from \$100 to \$120), in real terms (after considering the 10% inflation), the price has only increased by \$10 (from \$110 to \$120).

4. **Adjustment for Comparability**

   - Data from different time periods or sources can be hard to compare if there are inconsistencies in how they were collected or categorized.
   - **How to Adjust**:
     - *Standardize Units*: Make sure all data is in the same format (e.g., converting all sales figures to the same currency).
     - *Consistent Methods*: Use the same data collection methods over time to avoid biases.
   - **Example**: If one year sales are reported in thousands of dollars and the next year in millions, you must standardize the figures (e.g., convert millions to thousands) to accurately compare the two years.

12.2.9 Components of a Time Series

Time series data are typically decomposed into several components to better understand their underlying patterns. The first component is the **Trend component** (T_t), which represents the smooth, long-term direction of the time series.

DRAFT

Another key component is **Periodic Changes**, which include both **Seasonal** and **Cyclical components**. The **Seasonal component** (S_t) captures patterns that repeat within a year, such as monthly or quarterly fluctuations. In contrast, the **Cyclical component** (C_t) reflects longer-term rise and fall patterns that extend beyond a year.

Finally, the **Random component** or **Irregular Movements** (I_t) accounts for the residuals, which include unpredictable but identifiable episodic events and chance fluctuations that cannot be easily identified.

Component of Time Series: Time series decomposition divides data into

- trend
- periodic changes
  - seasonal
  - cyclical
- random components

to reveal underlying patterns and variations.

12.2.10 Methods of Time Series Analysis

1. **Decomposition:** Decomposition techniques break down a time series into its fundamental components:
   - **Classical Decomposition:**
     - Additive Model: This model assumes that the time series can be expressed as the sum of its components.
     - Multiplicative Model: This model assumes that the time series is the product of its components.
   - **STL (Seasonal and Trend decomposition using Loess):** A flexible method that can handle complex seasonal patterns and non-linear trends.
2. **Smoothing Methods:** Smoothing methods help to identify patterns by reducing noise:
   - **Moving Average:** Smooths a time series by averaging data points within a sliding window.
   - **Exponential Smoothing:**
     - Simple Exponential Smoothing

DRAFT

- Holt's Linear Trend Model
- Holt-Winters Seasonal Model

3. **Autoregressive Models:** These models use past values of a time series to predict future values:
    - AR (Autoregressive) Model
    - MA (Moving Average) Model
    - ARMA (Autoregressive Moving Average) Model
    - ARIMA (Autoregressive Integrated Moving Average) Model:
4. **Seasonal Models:** Models that account for seasonal patterns:
    - SARIMA (Seasonal ARIMA):
    - SARIMAX (Seasonal ARIMA with Exogenous Regressors)
5. **Exponential Smoothing State Space Model (ETS)**
6. **Machine Learning Methods:** Leverage machine learning algorithms to model time series data:
    - **Regression Trees and Random Forests**
    - **Support Vector Machines (SVM)**
    - **Neural Networks:** Includes LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Units).
7. **State Space Models:** Represent a time series as a system of equations:
    - **Kalman Filter:** A recursive algorithm for estimating the state of a linear dynamic system.
    - **Dynamic Linear Models (DLM):** Generalizes state space models to handle time-varying parameters.
8. **Fourier and Wavelet Transforms:** Analyze frequency components of a time series:
    - **Fourier Transform:** Decomposes a time series into sine and cosine components.
    - **Wavelet Transform:** Provides a time-frequency representation.
9. **Change Point Detection:** Identify points where the statistical properties of a time series change:
    - **CUSUM (Cumulative Sum Control Chart)**

DRAFT

- **Bayesian Change Point Analysis**

10. **Statistical Tests:** Check for specific properties in time series data:
    - **Stationarity Tests:** Such as the Augmented Dickey-Fuller (ADF) test or KPSS test.
    - **Seasonality Tests**
11. **Forecasting:** Predict future values based on historical data:
    - **ARIMA/SARIMA Forecasting**
    - **Exponential Smoothing Forecasting**
    - **Machine Learning Forecasting**

12.3 Main Components of a Time Series

A time series can be decomposed into several main components, each representing different aspects of the data. The primary components are:

- **Trend Component (T_t)**: This represents the long-term movement or direction in the data over time. The trend component shows whether the data is generally increasing, decreasing, or remaining constant over an extended period. Identifying the trend helps to understand the overall direction of the data.
- **Seasonal Component (S_t)**: This captures the regular, repeating patterns or cycles in the data that occur at specific intervals. For example, sales of certain products might peak during holidays every year, or temperature might follow a seasonal pattern. The seasonal component helps to account for these predictable fluctuations.
- **Cyclic Component (C_t)**: While similar to the seasonal component, the cyclic component refers to fluctuations that occur over longer, non-fixed periods, often influenced by economic or business cycles. Unlike seasonality, cycles are not fixed in length and can vary in duration. Identifying cycles helps to understand the impact of broader economic conditions on the data.
- **Irregular Component (I_t)**: Also known as the residual or error component, this represents the random, irregular variations that cannot be attributed to the trend, seasonal, or cyclic components. It includes noise or unexpected variations in the data. Analyzing the irregular component can help identify anomalies or outliers.

Understanding these components is essential for effective time series analysis, forecasting, and interpretation.

DRAFT

12.4 Mathematical Models for Time Series

Mathematical models for time series analysis are essential for understanding and forecasting time-dependent data. These models help capture patterns, trends, and relationships within the data. The main types of models are explained below.

12.4.1 Additive Model

The **additive model** is used to decompose a time series into its individual components. It assumes that the observed time series can be expressed as the sum of its components:

$$y_t = T_t + C_t + S_t + I_t \tag{12.1}$$

where y_t is the observed value at time t, T_t is the trend component, S_t is the seasonal component, C_t is the cyclic component, and I_t is the irregular component.

Limitations of the Additive Model

- Assumes that the magnitude of seasonal and cyclical fluctuations is constant, which may not always be true.
- May be too simplistic for more complex time series where seasonal and cyclical effects vary with the level of the data.

12.4.2 Multiplicative Model

The **multiplicative model** is used to decompose a time series into its various components. It assumes that the observed time series is the product of its components:

$$y_t = T_t \times C_t \times S_t \times I_t \tag{12.2}$$

Limitations of the Multiplicative Model

- Assumes that seasonal effects are proportional to the level of the time series, which may not always be the case.
- Can be complex to interpret due to the multiplicative nature of the model.

For the **AirPassengers** dataset, the multiplicative model is more appropriate because the seasonal variations in the data increase proportionally with the

DRAFT

level of the time series. As the number of passengers grows over time, the amplitude of seasonal fluctuations also increases, indicating that the seasonal component is relative to the trend. This proportional relationship aligns with the multiplicative model, which is suited for data where seasonal effects scale with the level of the series, unlike the additive model which assumes constant seasonal fluctuations. Therefore, the multiplicative model was used to decompose the **AirPassengers** time series presented in Figure 12.2, helping to isolate and analyze the various factors affecting the data.

12.4.3 Mixed Models

Mixed models (also known as **hybrid models**) combine aspects of both additive and multiplicative models to address complex time series data. They integrate features of both models to better capture the underlying patterns in the data.

A popular mixed model can be formulated as:

$$y_t = (T_t \times S_t \times C_t) + I_t.$$

Some other mixed models are

$$\left.\begin{aligned} y_t &= T_t C_t + S_t I_t \\ y_t &= T_t + C_t S_t I_t \\ y_t &= T_t + S_t + C_t I_t \end{aligned}\right\} \tag{12.3}$$

Limitations of Mixed Model

- Can become complex to specify and interpret.
- May require large amounts of data to accurately estimate all components.

s.

The model 12.2 or 12.2 can be use to obtain a measure of one or more of the components by elimination, viz., subtraction or division. For example, if trend component (T_t) is known, then using multiplication model, it can be isolated from the given time series to give:

$$S_t \times C_t \times I_t = \frac{y_t}{T_t} = \frac{\text{original values}}{\text{trend values}}$$

Thus, for the *annual data*, for which the seasonal component S_t is not here, we have

$$y_t = T_t \times C_t \times I_t \quad \Rightarrow \quad C_t \times I_t = \frac{y_t}{T_t}$$

DRAFT

Remark 12.4.1. *Time series data can be modeled using both additive and multiplicative components, depending on the nature of the data and the underlying patterns.*

- *Additive model is suitable when the seasonal variations and the trend are relatively constant over time. For example, if the seasonal fluctuations are roughly the same amount regardless of the level of the trend, an additive model might be appropriate.*
- *Multiplicative model is suitable when the seasonal variations and trend are proportional to the level of the series. For instance, if the amplitude of the seasonal fluctuations increases as the level of the series increases, a multiplicative model may be more appropriate.*

In practice, you can often determine which model fits better by examining the data, looking at plots, or applying statistical tests. Sometimes, a log transformation of a multiplicative model can simplify it into an additive form, which can then be analyzed using traditional additive methods.

12.5 Classical Decomposition

Classical decomposition is a technique used to separate a time series y_t into its constituent components: trend, seasonal, and residual. The method assumes that the time series can be expressed as a combination of these components, typically using either an additive or multiplicative model. Classical decomposition can be performed using either:

- **Additive Model**: $y_t = T_t + S_t + I_t$
- **Multiplicative Model**: $y_t = T_t \times S_t \times I_t$

In Python, the `sm.tsa.seasonal_decompose` function from the `statsmodels` library is commonly used for decomposing time series data. This function allows us to separate a time series into its components by using following methods.

- **Trend Component Estimation**:
  - The trend component is estimated using a **moving average** method.
  - The default window size for the moving average is based on the frequency of the time series. For instance, if the data is monthly, the default window size will typically be 12.
- **Seasonal Component Estimation**:
  - The seasonal component is computed as the average of the detrended data for each season (e.g., each month).

DRAFT

- This calculation considers the periodicity of the data, allowing it to capture repeating patterns effectively.

Note that the classical decomposition method (whether additive or multiplicative) does not include a separate component for cyclical variations. It only isolates trend, seasonal, and residual components. Cyclical variations might be present in the residual component if they exist, but classical decomposition does not explicitly model or separate these cycles. As a result, any cyclical patterns that are not captured by the trend or seasonal components are mixed into the residuals. To specifically analyze cyclical variations, other methods or models may be used, such as:

- Business Cycle Analysis: Specialized techniques for identifying and analyzing business cycles.
- Advanced Time Series Models: Models like SARIMA (Seasonal ARIMA) or state space models can account for more complex patterns, including cyclical variations.

Classical Decomposition Example for AirPassengers Dataset in Python

Below is an example of performing classical decomposition using the **AirPassengers** dataset. This code demonstrates how to decompose the time series into its trend, seasonal, and residual components using the `seasonal_decompose` function from the `statsmodels` library.

```
# Decompose the time series into trend, seasonal, and
    residual components
decomposition = seasonal_decompose(data, model='
    multiplicative')

# Extract components
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

# Plot the decomposed components
plt.figure(figsize=(12, 8))

plt.subplot(4, 1, 1)
plt.plot(data['Passengers'], label='Original Series')
plt.title('Original Series')
plt.legend(loc='best')

plt.subplot(4, 1, 2)
plt.plot(trend, label='Trend Component', color='orange')
plt.title('Trend Component')
```

DRAFT

```
plt.legend(loc='best')

plt.subplot(4, 1, 3)
plt.plot(seasonal, label='Seasonal Component', color='green')
plt.title('Seasonal Component')
plt.legend(loc='best')

plt.subplot(4, 1, 4)
plt.plot(residual, label='Residual Component', color='red')
plt.title('Residual Component')
plt.legend(loc='best')

plt.tight_layout()
plt.show()
```

Listing 12.2: Classical Decomposition of the AirPassengers Dataset

Figure 12.2 shows the decomposition of the **AirPassengers** Time Series into Trend, Seasonal, and Residual Components. The top subplot shows the original time series of monthly airline passenger numbers from 1949 to 1960. The original series exhibits a general upward trend, indicating an increase in passenger numbers over time. Additionally, there are recurring seasonal patterns, with higher passenger numbers in the summer months and lower numbers in the winter. The second subplot illustrates the trend component, representing the underlying long-term movement. The trend line indicates a steady increase in the number of passengers over the period. This long-term upward trajectory reflects the overall growth in the airline industry during these years. The third subplot displays the seasonal component, capturing the repeating annual patterns. The seasonal component shows regular peaks and troughs, highlighting the cyclical nature of the data. For example, higher passenger numbers in the summer months (June to August) and lower numbers in the winter months (December to February) are evident. The bottom subplot presents the residual component, which includes the remaining variability not explained by the trend or seasonal components. This plot shows random fluctuations and deviations from the trend and seasonal patterns. Large deviations might indicate unusual events or irregularities not captured by the model.

12.6 STL Decomposition

Seasonal and Trend decomposition using Loess (STL) is a robust method for decomposing a time series into its seasonal, trend, and residual components. It decomposition is more flexible than classical decomposition and offers several advantages, including handling cyclical variations more effectively. It utilizes locally weighted regression (loess) to smooth the data and extract these components, offering flexibility to handle complex and non-linear patterns. Uses loess smoothing to identify the long-term trend, adapting to local variations.

DRAFT

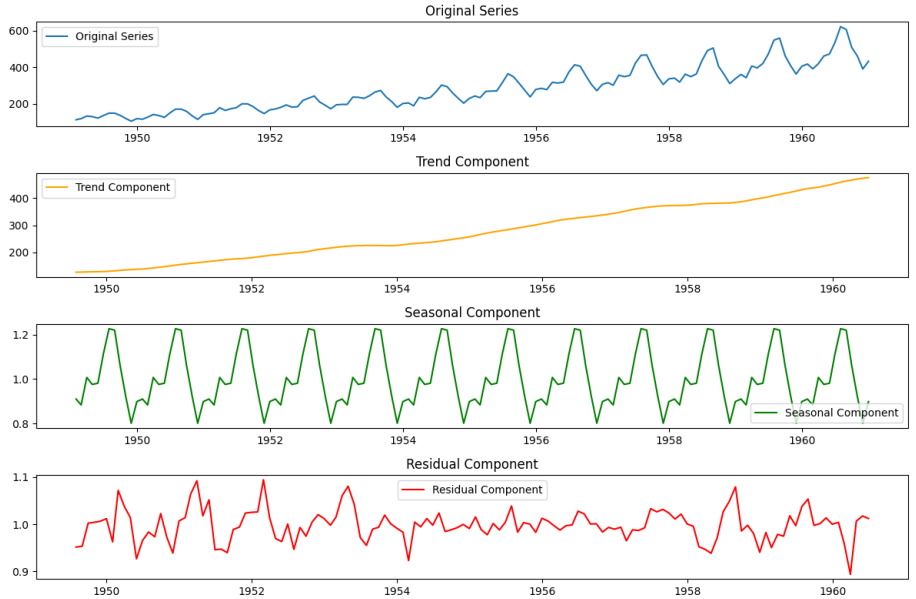


Figure 12.2: Classical decomposition of the **AirPassengers** Time Series into Trend, Seasonal, and Residual Components

Cyclical Patterns

Although STL does not explicitly decompose cyclical variations into a separate component, its method is well-suited to manage such patterns indirectly:

- **Residual Component**: The residual component in STL can include cyclical patterns if they are not captured by the trend or seasonal components. Since STL focuses on extracting smooth trend and seasonal components, any irregular cyclical behavior that does not fit neatly into these components might be reflected in the residuals.

- **Handling Irregular Cyclicality**: If cyclical behavior is present but not accounted for by the trend or seasonal extraction, it will likely be captured within the residual component, reflecting any remaining cyclical or irregular patterns in the data.

Advantages of STL over Classical Decomposition

- **Flexibility**: STL can adapt to changing seasonal patterns over time, unlike classical decomposition which assumes fixed seasonal effects. This adaptability can better capture complex cyclical variations.

- **Robustness**: The loess smoothing technique used in STL is robust to

DRAFT

outliers and irregularities in the data.

- **Visualization**: STL provides clear separation of trend, seasonal, and residual components, facilitating a better understanding of the time series.

STL Decomposition for AirPassengers Dataset in Python

Here is an example of performing STL decomposition on the **AirPassengers** dataset using Python. This code demonstrates how to decompose the time series into its trend, seasonal, and residual components, and plot the results.

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import STL

# Perform STL decomposition
stl = STL(data['Passengers'], seasonal=13)  # seasonal=13
    for monthly data with annual seasonality
result = stl.fit()

# Extract components
trend = result.trend
seasonal = result.seasonal
residual = result.resid

# Plot the decomposed components
plt.figure(figsize=(12, 8))

plt.subplot(4, 1, 1)
plt.plot(data['Passengers'], label='Original Series')
plt.title('Original Series')
plt.legend(loc='best')

plt.subplot(4, 1, 2)
plt.plot(trend, label='Trend Component', color='orange')
plt.title('Trend Component')
plt.legend(loc='best')

plt.subplot(4, 1, 3)
plt.plot(seasonal, label='Seasonal Component', color='green'
    )
plt.title('Seasonal Component')
plt.legend(loc='best')

plt.subplot(4, 1, 4)
plt.plot(residual, label='Residual Component', color='red')
plt.title('Residual Component')
```

```
plt.legend(loc='best')

plt.tight_layout()
plt.show()

```

Listing 12.3: STL Decomposition of the AirPassengers Dataset

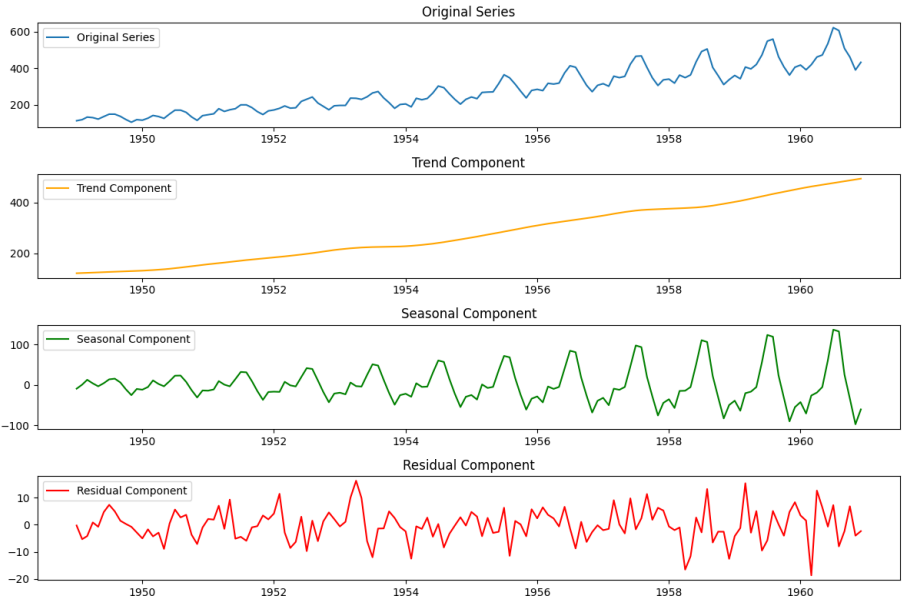


Figure 12.3: The STL Decomposition of the **AirPassengers** Time Series into Trend, Seasonal, and Residual Components

The Figure 12.3 illustrates the STL decomposition of the **AirPassengers** time series. The decomposition is separated into three main components:

- **Trend Component**: In the AirPassengers data, the trend shows a clear upward trajectory, indicating growth in passenger numbers over time.
- **Seasonal Component**: For the AirPassengers dataset, the seasonal component highlights the regular annual fluctuations, showing peaks and troughs corresponding to seasonal effects (e.g., higher passenger numbers during summer months).
- **Residual Component**: This represents the noise or irregularities that cannot be explained by the trend and seasonal components. It contains the leftover variation in the data after accounting for both trend and seasonal effects.

By analyzing these components, one can better understand the underlying patterns and make more informed predictions or decisions based on the time series data.

12.7 Measurement of Trend and Forecasting

The measurement of trend is essential in time series analysis as it allows for the identification and quantification of long-term movements within the data. Accurate trend measurement not only aids in understanding historical patterns but also serves as a foundation for forecasting future values. This section will discuss various techniques for measuring trends and the methodologies used in forecasting based on these trends. The following approaches can be used to measure trends:

1. Freehand or Graphic Method
2. Method of Semi-Averages
3. Method of Moving Averages
4. Method of Fitting Curve

12.7.1 Freehand or Graphic Method

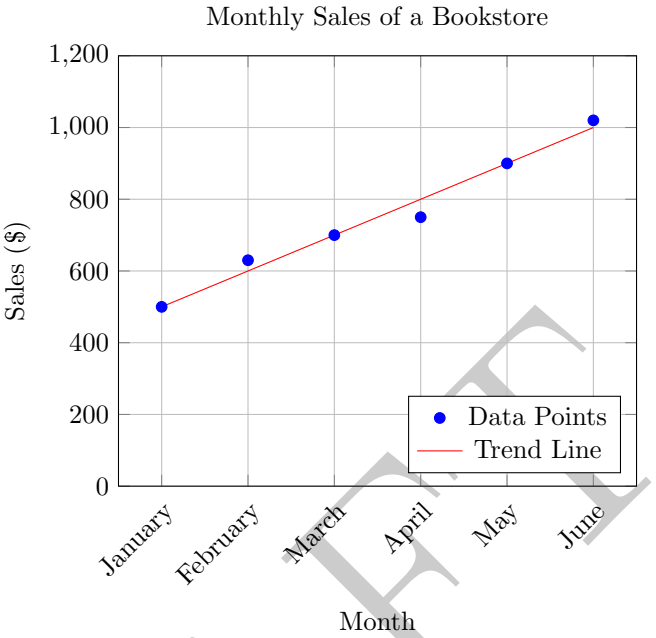
This method involves visually estimating the trend by plotting the data points on a graph and drawing a line that best represents the overall direction of the data. It is simple but can be subjective.

Example: Monthly Sales of a Bookstore

Let's say we are analyzing the monthly sales of a small bookstore over six months.

Month	Sales ($)
January	500
February	630
March	700
April	750
May	900
June	1020

DRAFT



After plotting the points, you can visually estimate the overall trend. The line drawn indicates an upward trend in sales. The trend line suggests that sales are consistently increasing over the months, indicating that the bookstore's sales performance is improving.

While this method is straightforward, it's essential to acknowledge that different individuals might draw slightly different lines based on their perception, leading to some subjectivity in the analysis.

12.7.2 Method of Semi-Averages

The Method of Semi-Averages is a statistical technique used to analyze trends in data by dividing the dataset into two equal parts, calculating the averages for each part, and then plotting these averages. This method helps to smooth out fluctuations and provide a clearer view of underlying trends.

Example: Monthly Sales of a Bookstore

Let's analyze the monthly sales data of a small bookstore over six months.

DRAFT

Month	Sales ($)
January	500
February	630
March	700
April	750
May	900
June	1020

Step 1: Split the Data
For the six-month period, we can divide the data into two halves:

- **First Half:** January, February, March
- **Second Half:** April, May, June

Step 2: Calculate Semi-Averages
- **First Half Average:**

$$\text{Average}_1 = \frac{500 + 630 + 700}{3} = \frac{1830}{3} = 610$$

- **Second Half Average:**

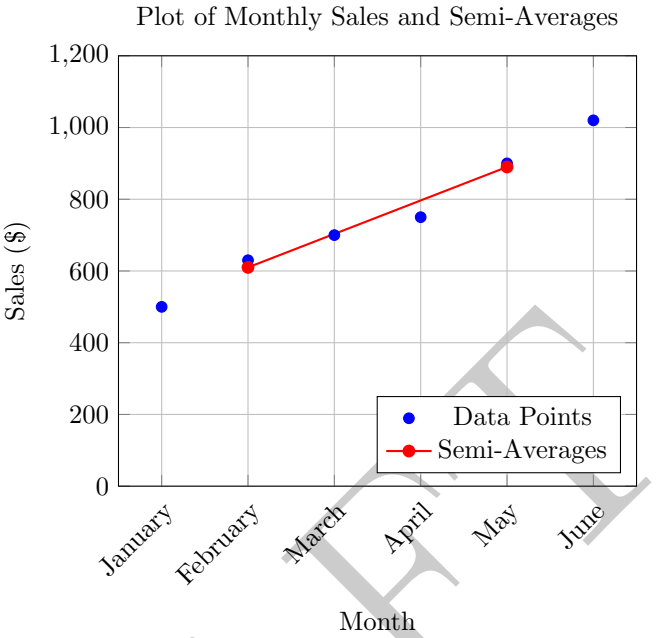
$$\text{Average}_2 = \frac{750 + 900 + 1020}{3} = \frac{2670}{3} = 890$$

Step 3: Summary of Semi-Averages
Now we can summarize the semi-averages in a table:

Time Period	Semi-Average ($)
First Half (Jan-Mar)	610
Second Half (Apr-Jun)	890

Step 4: Analyze the Trend
When plotted, the semi-averages show an upward trend from the first half (610) to the second half (890). This indicates that the bookstore's sales have increased over time, providing a clearer understanding of the overall trend without the noise of monthly fluctuations.

DRAFT



When applying the Method of Semi-Averages to datasets with an odd number of observations, the data can still be effectively analyzed by splitting it into two nearly equal parts. For example, if we have sales data for seven months, we can divide the dataset into a first half containing four observations and a second half with three. This division allows us to calculate semi-averages for each half, providing insights into trends over time. While the first half may be slightly more robust due to having one additional observation, the resulting averages still offer valuable information about changes in the dataset, highlighting trends and patterns despite the uneven split.

12.7.3 Method of Moving Averages

The Method of Moving Averages is a statistical technique used to analyze data trends by calculating averages over a specified number of consecutive observations. This approach helps to smooth out short-term fluctuations and highlight longer-term trends within the dataset. Moving averages are particularly useful in time series data, such as sales figures, to identify underlying patterns and inform decision-making.

The Method of Moving Averages provides a useful way to analyze trends in data by averaging out short-term variations. In this example, it effectively highlights the upward trend in bookstore sales, aiding in better decision-making regarding inventory and marketing strategies.

DRAFT

Types of Moving Averages

Here are the main types of moving averages:

- Simple Moving Average (SMA)
- Weighted Moving Average (WMA)
- Exponential Moving Average (EMA)
- Cumulative Moving Average (CMA)

Simple Moving Average (SMA)

The Simple Moving Average (SMA) is a method used to smooth out short-term fluctuations and highlight longer-term trends in time series data. It is calculated by taking the average of a fixed number of the most recent observations in a time series. The Simple Moving Average (SMA) at time t with a window size k is given by:

$$\text{SMA}_t = \frac{1}{k} \sum_{i=t-k+1}^{t} y_i$$

where:

- y_i is the observed value at time i,
- k is the number of periods over which the average is computed (window size),
- SMA_t is the SMA value at time t.

Example: Monthly Sales of a Bookstore

We have sales data for a bookstore over seven months:

Month	Sales ($)
January	500
February	630
March	700
April	750
May	900
June	1020
July	1100

DRAFT

Step 1: Choose a Window Size

For this example, we will use a **3-month moving average**, meaning we will calculate the average sales for every set of three consecutive months.

Step 2: Calculate the Moving Averages

The 3-month moving averages are calculated as follows:

- Moving Average for February, March, and April:

$$\text{MA}_{Feb-Mar-Apr} = \frac{500 + 630 + 700}{3} = \frac{1830}{3} = 610.00$$

- Moving Average for March, April, and May:

$$\text{MA}_{Mar-Apr-May} = \frac{630 + 700 + 750}{3} = \frac{2080}{3} \approx 693.33$$

- Moving Average for April, May, and June:

$$\text{MA}_{Apr-May-Jun} = \frac{700 + 750 + 900}{3} = \frac{2350}{3} \approx 783.33$$

- Moving Average for May, June, and July:

$$\text{MA}_{May-Jun-Jul} = \frac{750 + 900 + 1020}{3} = \frac{2670}{3} = 890.00$$

Step 3: Summary of Moving Averages

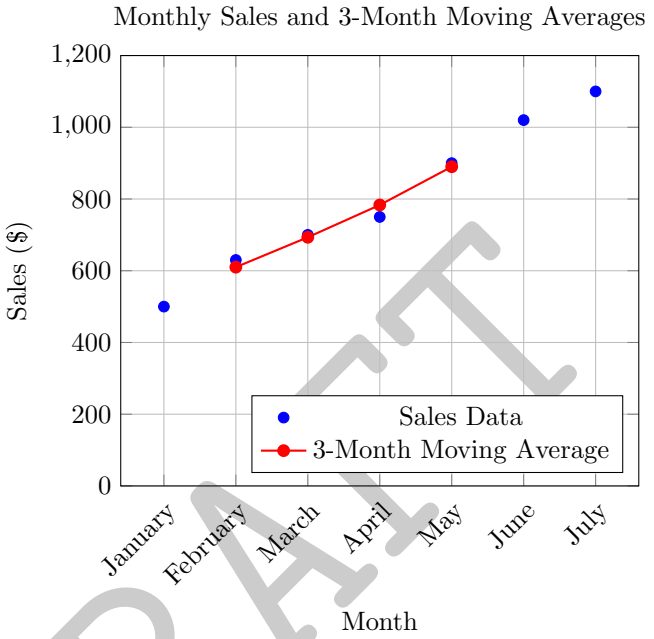
The calculated 3-month moving averages for the bookstore sales data are summarized in the following table:

Month	**Sales ($)**	**3-Month Moving Average ($)**
January	500	-
February	630	-
March	700	610.00
April	750	693.33
May	900	783.33
June	1020	890.00
July	1100	-

DRAFT

Step 4: Analyze the Trend

Plotting these moving averages alongside the original sales data allows us to visualize the trend. The moving averages reveal a gradual increase in sales, smoothing out fluctuations that may occur in the individual monthly data.



DRAFT

Python Code for SMA

The following Python code demonstrates how to calculate and visualize the Simple Moving Average (SMA) for the AirPassengers dataset described in Section 12.2.2. The code performs the following tasks:

- Load the AirPassengers dataset.
- Calculate the 12-month SMA.
- Plot the original data along with the SMA.

```
import pandas as pd
import matplotlib.pyplot as plt

# Load data
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/airline-passengers.csv'
data = pd.read_csv(url, parse_dates=['Month'], index_col='Month')
```

```
7
8   # Compute Simple Moving Average
9   data['SMA'] = data['Passengers'].rolling(window=12).mean()
10
11  # Plot the original data and SMA
12  plt.figure(figsize=(12, 6))
13  plt.plot(data['Passengers'], label='Original Data')
14  plt.plot(data['SMA'], label='12-Month SMA', color='red')
15  plt.title('Simple Moving Average')
16  plt.xlabel('Time')
17  plt.ylabel('Number of Passengers')
18  plt.legend()
19  plt.show()
```

Listing 12.4: Python Code to Calculate and Plot SMA for AirPassengers Dataset

The Simple Moving Average (SMA) of 12 months line of Air Passengers Data is presented in Figure 12.4.

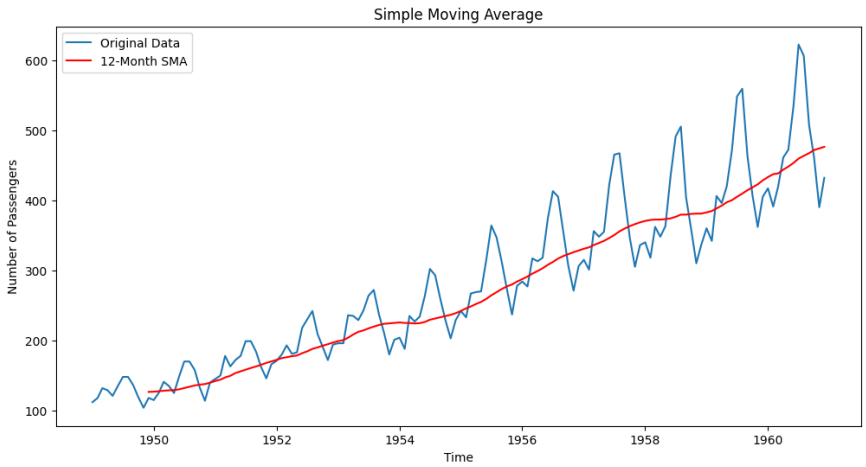


Figure 12.4: Simple Moving Average (SMA) of Air Passengers Data

Weighted Moving Average (WMA)

The Weighted Moving Average (WMA) is a technique used to smooth time series data by applying different weights to observations within the moving window. Unlike the Simple Moving Average (SMA), which gives equal weight to all observations, WMA assigns varying weights to different observations, allowing for more recent data to have a higher influence on the average. The Weighted Moving Average (WMA) at time t with weights $w_1, w_2, \ldots, w_k$ is given by:

$$\text{WMA}_t = \frac{\sum_{i=1}^{k} w_i \cdot y_{t-i+1}}{\sum_{i=1}^{k} w_i}$$

where:

- y_{t-i+1} is the observed value at time $t - i + 1$,
- w_i is the weight assigned to the i-th observation,
- k is the number of periods in the moving window,
- WMA_t is the weighted moving average value at time t.

WMA has several advantages over the SMA in time series analysis. Here are the key benefits:

- The WMA assigns different weights to data points within the window. Typically, more recent observations are given higher weights than older ones. This makes WMA more responsive to recent changes in the data, capturing trends more quickly. On the other hand, the Simple Moving Average gives equal weight to all data points within the window. This can result in a delayed response to recent changes because the contribution of past data points is just as significant as more recent ones.

- By giving more weight to recent observations, WMA can adapt more quickly to new trends or shifts in the time series. This can be particularly beneficial in rapidly changing environments where recent data is more relevant. On the other hand, SMA may smooth out trends more slowly because it averages all data points within the window equally, potentially missing out on more recent trends.

In essence, the Weighted Moving Average provides more flexibility and adaptability compared to the Simple Moving Average. It can better handle recent changes, reduce lag, and offer customization that can lead to more accurate and timely insights into the time series data.

Example: Monthly Sales of a Bookstore

We have sales data for a bookstore over seven months:

DRAFT

Month	Sales ($)
January	500
February	630
March	700
April	750
May	900
June	1020
July	1100

Step 1: Choose Weights

For this example, let's use the following weights for a 3-month WMA:

- Most recent month: 3
- Second most recent month: 2
- Third most recent month: 1

The total weight is $3 + 2 + 1 = 6$.

Step 2: Calculate the WMA

The Weighted Moving Averages for each applicable month are calculated as follows:

- WMA for March (January, February, March):

$$\text{WMA}_{Mar} = \frac{3 \times 700 + 2 \times 630 + 1 \times 500}{6} = \frac{3860}{6} \approx 643.33$$

- WMA for April (February, March, April):

$$\text{WMA}_{Apr} = \frac{3 \times 750 + 2 \times 700 + 1 \times 630}{6} = \frac{4280}{6} \approx 713.33$$

- WMA for May (March, April, May):

$$\text{WMA}_{May} = \frac{3 \times 900 + 2 \times 750 + 1 \times 700}{6} = \frac{4900}{6} \approx 816.67$$

- WMA for June (April, May, June):

$$\text{WMA}_{Jun} = \frac{3 \times 1020 + 2 \times 900 + 1 \times 750}{6} = \frac{5610}{6} \approx 935.00$$

DRAFT

- WMA for July (May, June, July):

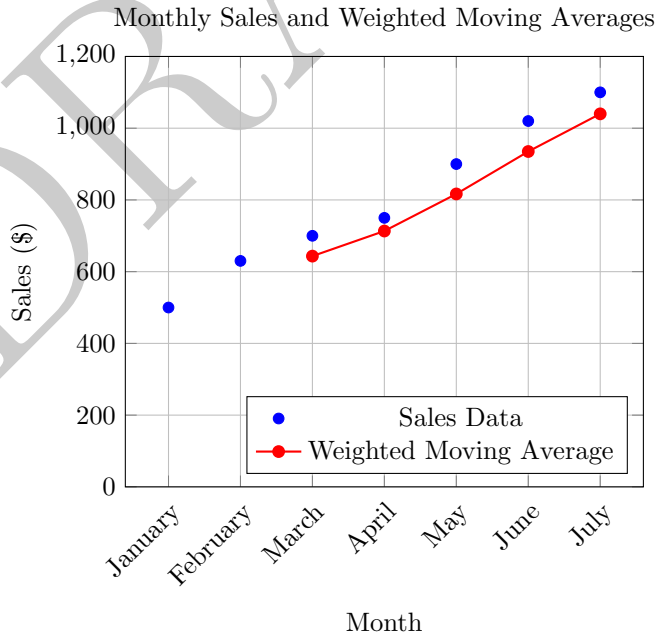
$$\text{WMA}_{Jul} = \frac{3 \times 1100 + 2 \times 1020 + 1 \times 900}{6} = \frac{6240}{6} = 1040.00$$

The calculated Weighted Moving Averages are summarized in the following table:

Month	Sales ($)	WMA ($)
January	500	-
February	630	-
March	700	643.33
April	750	713.33
May	900	816.67
June	1020	935.00
July	1100	1040.00

Step 3: Plotting Sales Data and WMA

The following plot shows the monthly sales data along with the calculated Weighted Moving Averages.



DRAFT

Python Code for WMA

The following Python code for 12-points weighted average is used for **AirPassengers** dataset described in Section 12.2.2.

```
import numpy as np

# Define weights
weights = np.arange(1, 13)

# Compute Weighted Moving Average
data['WMA'] = data['Passengers'].rolling(window=12).apply(
  lambda x: np.dot(weights, x) / weights.sum(), raw=True)

# Plot the original data and WMA
plt.figure(figsize=(12, 6))
plt.plot(data['Passengers'], label='Original Data')
plt.plot(data['WMA'], label='12-Month WMA', color='green')
plt.title('Weighted Moving Average')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()
```

Listing 12.5: Python Code to Calculate and Plot Weighted Moving Average (WMA)

The weighted moving average of 12 months line is presented in Figure 12.5.

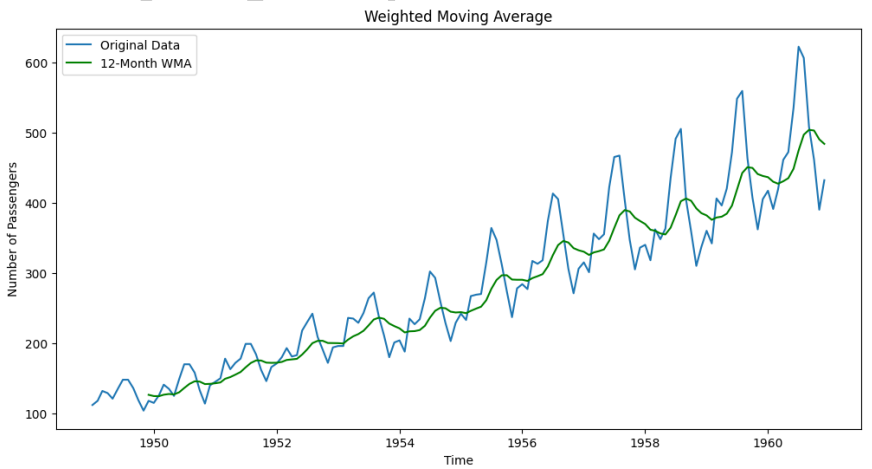


Figure 12.5: Weighted Moving Average of 12 Months.

12.7.4 Method of Fitting Curves

The Method of Fitting Curves is a statistical technique used to determine the best-fitting mathematical function that describes the relationship between two variables in a dataset. This method is commonly used in various fields, including economics, biology, and engineering, to analyze trends and make predictions based on empirical data.

The main purposes of fitting curves are:

- To find the underlying relationship between variables.
- To make predictions based on the fitted model.
- To understand the behavior of data over time or across different conditions.

Common Types of Curve Fitting

- **Linear Fitting:** In time series analysis, linear fitting involves estimating a straight-line trend through the data points over time. This method is used when the relationship between the variable and time appears to be constant.
  - **Equation:**
    $$y_t = mt + b$$
    where y_t represents the value at time t, m is the slope, and b is the y-intercept.
- **Polynomial Fitting:** This method involves fitting a polynomial equation to capture more complex trends in the data over time. It is useful for modeling non-linear relationships in time series data.
  - **General Form:**
    $$y_t = a_n t^n + a_{n-1} t^{n-1} + \ldots + a_1 t + a_0$$
    where $a_n, a_{n-1}, \ldots, a_0$ are coefficients, and t is the time variable.
- **Exponential Fitting:** Exponential fitting is used when data exhibits exponential growth or decay over time. This method models situations where the change in the variable is proportional to its current value.
  - **Equation:**
    $$y_t = ae^{bt}$$
    where a is the initial value and b is the growth (or decay) rate.

DRAFT

- **Logarithmic Fitting:** This approach is applied when data increases rapidly and then levels off. Logarithmic fitting can model phenomena with diminishing returns over time.

  - **Equation:**

    $$y_t = a \ln(t) + b$$

    where a and b are constants, and $\ln(t)$ is the natural logarithm of time.

- **Power Law Fitting:** Power law fitting is utilized to describe relationships where one variable scales as a power of another. This method is particularly effective in modeling heavy-tailed distributions often found in natural and social phenomena.

  - **Equation:**

    $$y_t = at^b$$

    where a is a constant and b is the exponent that characterizes the relationship.

Estimation of Trend Line Parameters

The trend model can be expressed as:

$$y_t = T_t + \varepsilon_t$$

In time series analysis, a simple linear equation estimates a linear trend over time. That is $T_t = \beta_0 + \beta_1 t$. Hence, a trend line can be expressed as

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

where y_t is the observed value, T_t is the trend component, β_0 is the intercept, β_1 is the slope, and ε_t is the error term.

Minimizing Function: The objective is to minimize the function:

$$\begin{aligned} Q = \min_{\beta_0,\beta_1} \sum_{t=1}^{n} \varepsilon_t^2 &= \min_{\beta_0,\beta_1} \sum_{t=1}^{n} (y_t - \widehat{y}_t)^2 \\ &= \min_{\beta_0,\beta_1} \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 t)^2 \end{aligned}$$

The estimators for β_0 and β_1 can be found by solving the following normalized equations:

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 t)^2 = 0 \tag{1}$$

DRAFT

and

$$\frac{\partial Q}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 t)^2 = 0 \tag{2}$$

From equation (1), we have:

$$\sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 t)(-1) = 0$$
$$\Rightarrow \sum_{t=1}^{n} y_t - n\beta_0 - \beta_1 \sum_{t=1}^{n} t = 0$$

This simplifies to:

$$\beta_0 = \bar{y} - \beta_1 \bar{t}$$

From equation (2), we have:

$$\sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 t)(-t) = 0$$
$$\Rightarrow \sum_{t=1}^{n} t y_t - \beta_0 \sum_{t=1}^{n} t - \beta_1 \sum_{t=1}^{n} t^2 = 0 \tag{3}$$

Substituting the value of β_0 into equation (3):

$$\sum_{t=1}^{n} t y_t - (\bar{y} - \beta_1 \bar{t}) \sum_{t=1}^{n} t - \beta_1 \sum_{t=1}^{n} t^2 = 0$$
$$\Rightarrow \sum_{t=1}^{n} t y_t - \bar{y} \sum_{t=1}^{n} t + \beta_1 \bar{t} \sum_{t=1}^{n} t - \beta_1 \sum_{t=1}^{n} t^2 = 0$$
$$\Rightarrow \sum_{t=1}^{n} t y_t - n\bar{y}\bar{t} + n\beta_1 \bar{t}^2 - \beta_1 \sum_{t=1}^{n} t^2 = 0$$
$$\Rightarrow \sum_{t=1}^{n} t y_t - n\bar{y}\bar{t} - \beta_1 \left(\sum_{t=1}^{n} t^2 - n\bar{t}^2 \right) = 0$$

$$\Rightarrow \beta_1 = \frac{\sum_{t=1}^{n} t y_t - n\bar{t}\bar{y}}{\sum_{t=1}^{n} t^2 - n\bar{t}^2}$$

Hence, the least squares estimators of β_0 and β_1 are:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{t}$$

$$\widehat{\beta_1} = \frac{\sum_{t=1}^{n} t y_t - n\bar{t}\bar{y}}{\sum_{t=1}^{n} t^2 - n\bar{t}^2}$$

DRAFT

The expression for $\widehat{\beta_1}$ can also be rewritten as:

$$\widehat{\beta_1} = \frac{\sum_{t=1}^{n}(t-\bar{t})(y_t-\bar{y})}{\sum_{t=1}^{n}(t-\bar{t})^2}$$

because

$$\sum_{t=1}^{n}(t-\bar{t})^2 = \sum_{t=1}^{n}t^2 - n\bar{t}^2 \quad \text{and} \quad \sum_{t=1}^{n}(t-\bar{t})(y_t-\bar{y}) = \sum_{t=1}^{n}ty_t - n\bar{t}\bar{y}$$

Example: Consider the following data representing annual sales (in thousands dollar) over five years:

Year (t)	Sales (y_t)
2018	7
2019	10
2020	9
2021	11
2022	13

Consider the following dataset representing sales over five years:

Year (t)	Sales (y_t)
1	7
2	10
3	9
4	11
5	13

We have,

$$\bar{y} = \frac{7+10+9+11+13}{5} = 10$$

$$\bar{t} = \frac{1+2+3+4+5}{5} = 3$$

Calculating the necessary sums in the following table.

DRAFT

Year (t)	Sales (y_t)	$t \cdot y_t$	t^2
1	7	7	1
2	10	20	4
3	9	27	9
4	11	44	16
5	13	65	25
Total		163	55

Now substituting into the formula for $\widehat{\beta_1}$:

$$\widehat{\beta_1} = \frac{\sum_{t=1}^{5} t y_t - n\bar{t}\bar{y}}{\sum_{t=1}^{5} t^2 - n\bar{t}^2} = \frac{163 - 5 \times 3 \times 10}{55 - 5 \times 55} = 1.3$$

Now, calculate $\widehat{\beta_0}$:

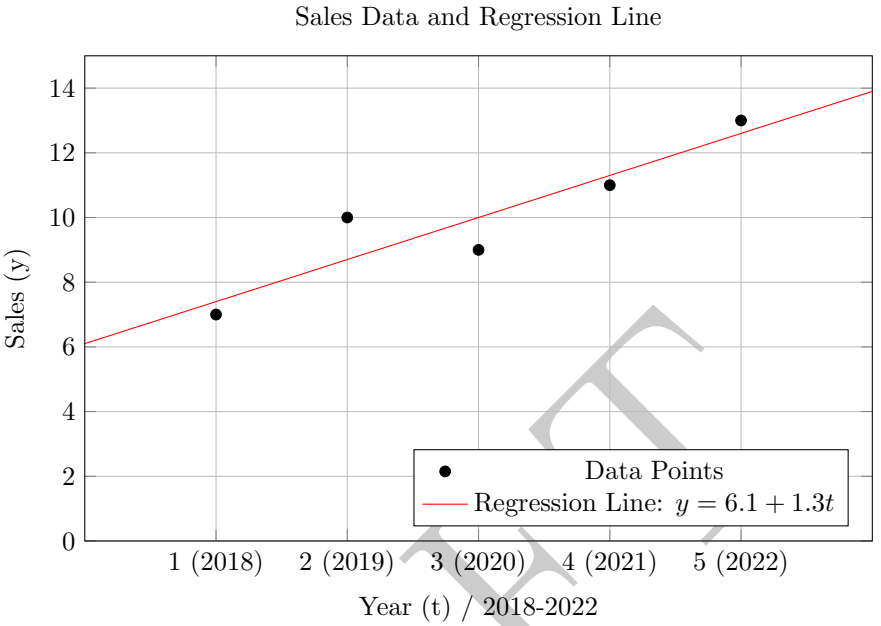
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{t} = 10 - 1.3 \times 3 = 6.1$$

Thus, the estimated linear regression equation is:

$$y_t = 6.1 + 1.3t$$

This means that for each additional year, sales increase by an average of 1.3 thousand dollars, starting from an intercept of 6.1 thousand dollars when $t = 0$. Now, we will plot the data points and the regression line.

DRAFT



Polynomial Equation

Definition: Polynomial regression extends linear regression to model more complex trends in time series data. The model can be expressed as:

$$y_t = a_n t^n + a_{n-1} t^{n-1} + \ldots + a_1 t + a_0 + \epsilon_t$$

Example: Consider the following data representing annual average temperatures (in °C) over four years:

Year (t)	Temperature (y_t)
2018	30
2019	31
2020	33
2021	36

A quadratic fit may yield:

$$y_t = 28 + 1.5(t - 2018) + 0.5(t - 2018)^2$$

Spline Fitting Curve

Definition: Spline fitting in time series involves using piecewise polynomial functions to model trends.

DRAFT

Example: Consider the following data representing annual stock prices (in USD) over six years:

Year (t)	Price (y_t)
2016	100
2017	110
2018	115
2019	130
2020	125
2021	140

Using cubic splines, you can create a smooth curve through these points.

Exponential Fitting Curve

Definition: Exponential fitting models trends that grow or decay at a constant percentage rate.

Example: Consider the following data representing annual population growth (in thousands) over five years:

Year (t)	Population (y_t)
2018	1000
2019	1100
2020	1200
2021	1400
2022	1800

An exponential fit might yield:

$$y_t = 1000e^{0.15(t-2018)}$$

Logarithmic Fitting Curve

Definition: Logarithmic fitting is used for data that grows quickly and then levels off.

Example: Consider the following data representing learning performance scores over several years:

DRAFT

Year (t)	Score (y_t)
2018	10
2019	20
2020	25
2021	28
2022	30

A logarithmic fit might yield:

$$y_t = 5 + 15 \ln(t - 2017)$$

Power Law Fitting Curve

Definition: Power law fitting describes relationships where one variable varies as a power of another.

Example: Consider the following data representing the distribution of city populations (in thousands):

City Rank (t)	Population (y_t)
1	500
2	250
3	125
4	60
5	30

A power law fit might yield:

$$y_t = 600t^{-1.5}$$

Choosing the appropriate method for trend estimation in time series analysis depends on the nature of the data, the expected trend complexity, and the need for interpretability versus accuracy in forecasting.

Detrending

Detrending is a statistical technique used to remove trends from time series data. This process allows for a clearer analysis of the data's underlying patterns, such as seasonal fluctuations and cyclical variations. In an additive model, the time series can be expressed as:

DRAFT

$$Y_t = T_t + S_t + C_t + I_t$$

To **detrend** an additive time series, you can subtract the trend component from the original data:

$$D_t = Y_t - T_t$$

The resulting values D_t will consist of the seasonal, cyclical and irregular components.

In a multiplicative model, the time series is expressed as:

$$Y_t = T_t \times S_t \times C_t \times I_t$$

To **detrend** a multiplicative time series,we divide the original data by the trend component:

$$D_t = \frac{Y_t}{T_t}$$

The resulting values D_t will be a product of the seasonal, cyclical and irregular components.

12.8 Measurement of Seasonal Variation

Seasonal variation refers to regular and predictable changes that recur in a time series data over specific periods, such as days, months, or quarters. Understanding and measuring these variations is crucial for accurate forecasting and analysis.

Importance of Measuring Seasonal Variation

Measuring seasonal variation helps analysts to:

- Identify patterns and trends within the data.
- Make informed decisions based on historical seasonal performance.
- Improve forecasting accuracy by accounting for seasonal effects.

Several methods can be employed to measure seasonal variation, including:

(i). Method of Simple Average

(ii). Ratio-to-Trend Method

(iii). Ratio-to-Moving-Average Method

(iv). Method of Link Relatives

DRAFT

12.8.1 Method of Simple Averages

This method involves calculating the average for each season (or quarter) over a specified number of years and then comparing these averages to the overall average.

Steps

1. Calculate the average sales for each season:

$$\bar{y}_i = \frac{1}{k} \sum_{j=1}^{k} y_{ij}$$

   where $\bar{y}_i$ is the seasonal average for season i, y_{ij} is the sales for year j in season i, and k is the number of years.

2. Calculate the overall average:

$$\text{Overall Average } \bar{y} = \frac{1}{m} \sum_{i=1}^{m} \bar{y}_i$$

   where m is the number of seasons.

3. Calculate the seasonal indices:

$$\text{Seasonal Index}_i = \frac{\bar{y}_i}{\text{Overall Average}} \times 100 = \frac{\bar{y}_i}{\bar{y}} \times 100$$

Let us consider following problems to explain this method.

Problem 12.1. *The following table presents the quarterly sales data over four years:*

Year	Q1	Q2	Q3	Q4
2021	30	40	36	34
2022	34	52	50	44
2023	40	58	54	48
2024	80	92	86	82

Table 12.1: Quarterly Sales Data

Using simple average method, determine the quarterly seasonal indices for the adjoining data.

DRAFT

Solution

- Seasonal averages:

$$\bar{y}_1 = \frac{30 + 34 + 40 + 80}{4} = 46$$

$$\bar{y}_2 = \frac{40 + 52 + 58 + 92}{4} = 60.5$$

$$\bar{y}_3 = \frac{36 + 50 + 54 + 86}{4} = 56.5$$

$$\bar{y}_4 = \frac{34 + 44 + 48 + 82}{4} = 52$$

- Overall average:

$$\text{Overall Average: } \bar{y} = \frac{30 + 40 + 36 + 34 + \cdots + 92 + 86 + 82}{16} = 54.5$$

- Seasonal indices:

$$\text{Seasonal Index}_{Q1} = \frac{46}{54.5} \times 100 \approx 84.4$$

$$\text{Seasonal Index}_{Q2} = \frac{60.5}{54.5} \times 100 \approx 110.9$$

$$\text{Seasonal Index}_{Q3} = \frac{56.5}{54.5} \times 100 \approx 103.7$$

$$\text{Seasonal Index}_{Q4} = \frac{52}{54.5} \times 100 \approx 95.4$$

Problem 12.2. *Determine the monthly seasonal indices for the following data of production of a commodity for the years 2010, 2011, and 2012 using the method of simple averages.*

DRAFT

Table 12.2: Production data in Tonnes

Months	**2010**	**2011**	**2012**
January	120	150	160
February	110	140	150
March	100	130	140
April	140	160	160
May	150	160	150
June	150	150	170
July	160	170	160
August	130	120	130
September	110	1360	100
October	100	120	100
November	120	130	110
December	150	140	150

Solution

First of all, we arrange the data as shown in columns 1 to 4 of the Table 12.3 given below: We are given the monthly production of a commodity for 3 years. For calculating the monthly seasonal indices, we first calculate the month-wise total production for the 3 years. Then we calculate the monthly averages for all 12 months. Note from Table 12.3 that for January, it is 143.3, for February, it is 133.3, ..., etc. Next, we calculate the average of all monthly averages, i.e.

$$\text{Overall Average: } \bar{y} = \frac{143.3 + 133.3 + \ldots + 146.6}{12} = 136.625$$

Now we calculate the seasonal indices by taking the percentage of monthly averages $\bar{y}_i$ to the combined average $\bar{y}$, one at a time, for $i = 1, 2, \ldots, 12$:

DRAFT

Table 12.3: Seasonal indices for the given time series

Months	**2010**	**2011**	**2012**	**Monthly Avgs.**	**Seasonal Index**
January	120	150	160	143.3	104.886
February	110	140	150	133.3	97.566
March	100	130	140	123.3	90.247
April	140	160	160	153.3	112.205
May	150	160	150	153.3	112.205
June	150	150	170	156.6	114.620
July	160	170	160	163.3	119.524
August	130	120	130	126.6	92.662
September	100	130	100	113.3	82.928
October	100	120	1000	106.6	78.024
November	120	130	110	120.0	87.832
December	150	140	150	146.6	107.301
Total	-	-	-	1639.5	1200
Average	-	-	-	136.625	100

$$\text{Seasonal Index for January} = \frac{143.3}{136.625} \times 100 = 104.886$$
$$\text{Seasonal Index for February} = \frac{133.3}{136.625} \times 100 = 97.566$$
$$\text{Seasonal Index for March} = \frac{123.3}{136.625} \times 100 = 90.247$$
$$\text{Seasonal Index for April} = \frac{153.3}{136.625} \times 100 = 112.205$$
$$\text{Seasonal Index for May} = \frac{153.3}{136.625} \times 100 = 112.205$$
$$\text{Seasonal Index for June} = \frac{156.6}{136.625} \times 100 = 114.620$$
$$\text{Seasonal Index for July} = \frac{163.3}{136.625} \times 100 = 119.524$$
$$\text{Seasonal Index for August} = \frac{126.6}{136.625} \times 100 = 92.662$$
$$\text{Seasonal Index for September} = \frac{113.3}{136.625} \times 100 = 82.928$$
$$\text{Seasonal Index for October} = \frac{106.6}{136.625} \times 100 = 78.024$$
$$\text{Seasonal Index for November} = \frac{120}{136.625} \times 100 = 87.832$$
$$\text{Seasonal Index for December} = \frac{146.6}{136.625} \times 100 = 107.301$$

DRAFT

Advantages of Method of Simple Averages

1. This method allows for rapid computation of seasonal indices, making it useful for preliminary analyses or when time is of the essence.

2. The results obtained from this method are easy to interpret, which is beneficial for communicating findings to stakeholders who may not have a statistical background.

Disadvantages of Method of Simple Averages

1. Neglect of Trends and Cyclical Components: It assumes that the data do not contain any trends or cyclical components, which is often not the case in real-world economic and business time series.

2. Limited Practical Utility: Since many economic and business time series exhibit trends and cycles, the Method of Simple Averages, while easy to apply, may lack practical utility in providing meaningful insights in these contexts.

3. Inability to Capture Seasonal Fluctuations: This method may not effectively capture seasonal variations, especially if they are influenced by underlying trends or cycles.

4. By relying solely on averages, this method may oversimplify the complexity of the data, leading to inaccurate conclusions.

12.8.2 Ratio to Trend Method

This method helps calculate seasonal indices without the influence of trends, improving upon the simple average method. Here's how it works in easy steps:

1. **Calculate Trend Values**: Use a method like least squares to find trend values for each time period. This can involve fitting a straight line or a curve to the data.

2. **Calculate Ratios**: For each original value, divide it by the corresponding trend value, then multiply by 100. This gives you percentage values that are free from trends.

3. **Find Seasonal Indices**: To eliminate any irregular fluctuations, calculate the average (or median) of these percentage values for each season across the years. If there are extreme values, it's better to use the median to avoid skewing results. If there are only a few unusual values, use the mean instead.

4. **Check Total Indices**: If you are looking at quarterly data, the sum of the seasonal indices should equal 400. For monthly data, it should equal 1200. Sometimes, the calculated total won't be exact, but that's okay.

DRAFT

By following the step-by-step procedure outlined above, we can compute the seasonal indices using this method. This approach is grounded in solid logic and leverages comprehensive data.

Problem 12.3. *Compute the seasonal indices for the following time series of sales (in thousand £) of a commodity by the ratio to trend method:*

Table 12.4: Sales data (in thousand £) of a commodity

Year	I Quarter	II Quarter	III Quarter	IV Quarter
2008	800	920	880	820
2009	540	760	680	620
2010	400	580	540	480
2011	340	520	500	440
2012	300	400	360	340

Solution

In analyzing the sales data over the specified quarters, we utilized regression analysis to derive a predictive model. The results indicate that the best-fit line for the data can be expressed by the equation (using the following Python code):

$$\widehat{y}_t = 860.84 - 28.56t$$

Here, we consider t values for each quarter from 1 to 20, along with the corresponding y values (sales). This equation suggests that as time progresses (represented by t), the expected sales ($\widehat{y}_t$) decrease by approximately 28.56 units for each quarter, starting from an intercept of 860.84. This model highlights the declining trend in sales, providing valuable insights for forecasting future performance and making informed business decisions.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

# Prepare the data
time_index = np.array(range(1, 21)).reshape(-1, 1)
sales = np.array([800, 920, 880, 820, 540, 760, 680, 620,
    400, 580, 540, 480, 340, 520, 500, 440, 300, 400, 360,
    340])

# Fit the model
model = LinearRegression()
```

DRAFT

```
model.fit(time_index, sales)

# Get the slope and intercept
slope = model.coef_[0]
intercept = model.intercept_

# Predict sales using the linear model
predicted_sales = model.predict(time_index)

# Plotting the data and the fitted line
plt.scatter(time_index, sales, color='blue', label='Actual Sales Data')
plt.plot(time_index, predicted_sales, color='red', label='Fitted Linear Line')
plt.xlabel('Time Index')
plt.ylabel('Sales (in thousand)')
plt.title('Linear Fit of Sales Data')
plt.legend()
plt.show()

print(f'Slope: {slope}, Intercept: {intercept}')
```

Listing 12.6: Code for finding trend line.

Step 1: Calculate Trend Values

Using the trend equation, we compute the trend values for each quarter from 2008 to 2012.

DRAFT

$$\widehat{y}_1 = 860.84 - 28.56 \cdot 1 = 832.28 \quad \text{(Q1 2008)}$$
$$\widehat{y}_2 = 860.84 - 28.56 \cdot 2 = 803.72 \quad \text{(Q2 2008)}$$
$$\widehat{y}_3 = 860.84 - 28.56 \cdot 3 = 775.16 \quad \text{(Q3 2008)}$$
$$\widehat{y}_4 = 860.84 - 28.56 \cdot 4 = 746.60 \quad \text{(Q4 2008)}$$
$$\widehat{y}_5 = 860.84 - 28.56 \cdot 5 = 718.04 \quad \text{(Q1 2009)}$$
$$\widehat{y}_6 = 860.84 - 28.56 \cdot 6 = 689.48 \quad \text{(Q2 2009)}$$
$$\widehat{y}_7 = 860.84 - 28.56 \cdot 7 = 660.92 \quad \text{(Q3 2009)}$$
$$\widehat{y}_8 = 860.84 - 28.56 \cdot 8 = 632.36 \quad \text{(Q4 2009)}$$
$$\widehat{y}_9 = 860.84 - 28.56 \cdot 9 = 603.80 \quad \text{(Q1 2010)}$$
$$\widehat{y}_{10} = 860.84 - 28.56 \cdot 10 = 575.24 \quad \text{(Q2 2010)}$$
$$\widehat{y}_{11} = 860.84 - 28.56 \cdot 11 = 546.68 \quad \text{(Q3 2010)}$$
$$\widehat{y}_{12} = 860.84 - 28.56 \cdot 12 = 518.12 \quad \text{(Q4 2010)}$$
$$\widehat{y}_{13} = 860.84 - 28.56 \cdot 13 = 489.56 \quad \text{(Q1 2011)}$$
$$\widehat{y}_{14} = 860.84 - 28.56 \cdot 14 = 461.00 \quad \text{(Q2 2011)}$$
$$\widehat{y}_{15} = 860.84 - 28.56 \cdot 15 = 432.44 \quad \text{(Q3 2011)}$$
$$\widehat{y}_{16} = 860.84 - 28.56 \cdot 16 = 403.88 \quad \text{(Q4 2011)}$$
$$\widehat{y}_{17} = 860.84 - 28.56 \cdot 17 = 375.32 \quad \text{(Q1 2012)}$$
$$\widehat{y}_{18} = 860.84 - 28.56 \cdot 18 = 346.76 \quad \text{(Q2 2012)}$$
$$\widehat{y}_{19} = 860.84 - 28.56 \cdot 19 = 318.20 \quad \text{(Q3 2012)}$$
$$\widehat{y}_{20} = 860.84 - 28.56 \cdot 20 = 289.64 \quad \text{(Q4 2012)}$$

All trend values are tabulated in Table 12.5.

Table 12.5: Trend Values

Year	**Q1**	**Q2**	**Q3**	**Q4**
2008	832.28	803.72	775.16	746.60
2009	718.04	689.48	660.92	632.36
2010	603.80	575.24	546.68	518.12
2011	489.56	461.00	432.44	403.88
2012	375.32	346.76	318.20	289.64

Step 2: Calculate Ratios

Next, we compute the ratios of actual sales to trend values for each quarter using the formula:

DRAFT

$$\text{Ratio} = \frac{\text{Actual Sales}}{\text{Trend Value}} \times 100$$

Q1 Ratios

$$\text{Q1 2008} : \frac{800}{832.28} \times 100 \approx 96.15$$
$$\text{Q1 2009} : \frac{540}{718.04} \times 100 \approx 75.16$$
$$\text{Q1 2010} : \frac{400}{603.80} \times 100 \approx 66.22$$
$$\text{Q1 2011} : \frac{340}{489.56} \times 100 \approx 69.49$$
$$\text{Q1 2012} : \frac{300}{375.32} \times 100 \approx 79.96$$

Q2 Ratios

$$\text{Q2 2008} : \frac{920}{803.72} \times 100 \approx 114.48$$
$$\text{Q2 2009} : \frac{760}{689.48} \times 100 \approx 110.24$$
$$\text{Q2 2010} : \frac{580}{575.24} \times 100 \approx 100.83$$
$$\text{Q2 2011} : \frac{520}{461.00} \times 100 \approx 112.76$$
$$\text{Q2 2012} : \frac{400}{346.76} \times 100 \approx 115.25$$

Q3 Ratios

$$\text{Q3 2008} : \frac{880}{775.16} \times 100 \approx 113.52$$
$$\text{Q3 2009} : \frac{680}{660.92} \times 100 \approx 102.74$$
$$\text{Q3 2010} : \frac{540}{546.68} \times 100 \approx 98.77$$
$$\text{Q3 2011} : \frac{500}{432.44} \times 100 \approx 115.62$$
$$\text{Q3 2012} : \frac{360}{318.20} \times 100 \approx 113.13$$

DRAFT

Q4 Ratios

$$\text{Q4 2008} : \frac{820}{746.60} \times 100 \approx 109.80$$
$$\text{Q4 2009} : \frac{620}{632.36} \times 100 \approx 97.96$$
$$\text{Q4 2010} : \frac{480}{518.12} \times 100 \approx 92.68$$
$$\text{Q4 2011} : \frac{440}{403.88} \times 100 \approx 108.95$$
$$\text{Q4 2012} : \frac{340}{289.64} \times 100 \approx 117.16$$

The results are tabulated in Table 12.6.

Step 3: Compute Seasonal Indices

We eliminate any irregular fluctuations and calculate the average (or median) of these percentage values for each season across the years. Here, we consider the average of these ratios in each quarter for different years. The results are presented in Table 12.6.

Table 12.6: Summary of Ratios (%) and Adjusted Seasonal Indices for Each Quarter

Year	**Q1**	**Q2**	**Q3**	**Q4**
2008	96.15	114.48	113.52	109.80
2009	75.16	110.24	102.74	97.96
2010	66.22	100.83	98.77	92.68
2011	69.49	112.76	115.62	108.95
2012	79.96	115.25	113.13	117.16
Average (Seasonal Index)	**77.40**	**110.71**	**108.76**	**105.31**
Adjusted Seasonal Index	**76.97**	**110.12**	**108.18**	**104.73**

Step 4: Adjustment of Seasonal Indices

Since we are working with quarterly data, the sum of the seasonal indices should ideally be 400. The initial seasonal indices calculated were as follows:

- Q1 Seasonal Index = 77.40
- Q2 Seasonal Index = 110.71

DRAFT

- Q3 Seasonal Index = 108.76
- Q4 Seasonal Index = 105.31

The total sum of these indices was:

$$77.40 + 110.71 + 108.76 + 105.31 = 402.18$$

Since the total is greater than 400, we apply an adjustment factor to scale the indices down proportionally:

$$\text{Adjustment Factor} = \frac{400}{402.18} \approx 0.99457$$

Multiplying each seasonal index by this adjustment factor, we obtain the adjusted seasonal indices:

- Adjusted Q1 Seasonal Index = $77.40 \times 0.99457 \approx 76.97$
- Adjusted Q2 Seasonal Index = $110.71 \times 0.99457 \approx 110.12$
- Adjusted Q3 Seasonal Index = $108.76 \times 0.99457 \approx 108.18$
- Adjusted Q4 Seasonal Index = $105.31 \times 0.99457 \approx 104.73$

The total of the adjusted seasonal indices is:

$$76.97 + 110.12 + 108.18 + 104.73 = 400.00$$

Thus, the seasonal indices now sum to 400, as required. Below is the final table of ratios and adjusted seasonal indices.

Table 12.7: Summary of Ratios (%) and Adjusted Seasonal Indices for Each Quarter

Year	**Q1**	**Q2**	**Q3**	**Q4**
2008	96.15	114.48	113.52	109.80
2009	75.16	110.24	102.74	97.96
2010	66.22	100.83	98.77	92.68
2011	69.49	112.76	115.62	108.95
2012	79.96	115.25	113.13	117.16
Average (Seasonal Index)	**77.40**	**110.71**	**108.76**	**105.31**
Adjusted Seasonal Index	**76.97**	**110.12**	**108.18**	**104.73**

Finally, we compute the seasonal indices by averaging the ratios for each quarter. These indices represent the seasonal variation in sales for each quarter relative to the trend values calculated from the sales data.

DRAFT

Forecasting Sales for Next Year

To forecast the sales for the next year using the seasonal indices and trend values, we follow these steps:

Trend Equation

The trend equation calculated from the historical data is:

$$\widehat{y}_t = 860.84 - 28.56t$$

where $\widehat{y}_t$ is the estimated sales (trend component) for time period t, and t represents the time periods (1 to 20 for historical data).

Forecasting for Next Year

For 2013, we will forecast the sales for the next four quarters, corresponding to time periods $t = 21$ to $t = 24$. We compute the trend values for the next four quarters using the trend equation:

$$\begin{aligned}
&\text{For Q1 } (t = 21): \quad \widehat{y}_{21} = 860.84 - 28.56 \times 21 = 261.08 \\
&\text{For Q2 } (t = 22): \quad \widehat{y}_{22} = 860.84 - 28.56 \times 22 = 232.52 \\
&\text{For Q3 } (t = 23): \quad \widehat{y}_{23} = 860.84 - 28.56 \times 23 = 203.96 \\
&\text{For Q4 } (t = 24): \quad \widehat{y}_{24} = 860.84 - 28.56 \times 24 = 175.40
\end{aligned}$$

We use the seasonal indices to adjust the trend values. The seasonal indices (adjusted) are:

$$\text{Q1: } 76.97 \quad \text{Q2: } 110.12 \quad \text{Q3: } 108.18 \quad \text{Q4: } 104.73$$

We multiply each trend value by the corresponding seasonal index (converted to a factor by dividing by 100):

$$\begin{aligned}
&\text{For Q1:} \quad \text{Forecast} = 261.08 \times \frac{76.97}{100} = 261.08 \times 0.7697 \approx 200.97 \\
&\text{For Q2:} \quad \text{Forecast} = 232.52 \times \frac{110.12}{100} = 232.52 \times 1.1012 \approx 256.07 \\
&\text{For Q3:} \quad \text{Forecast} = 203.96 \times \frac{108.18}{100} = 203.96 \times 1.0818 \approx 220.49 \\
&\text{For Q4:} \quad \text{Forecast} = 175.40 \times \frac{104.73}{100} = 175.40 \times 1.0473 \approx 183.63
\end{aligned}$$

Hence, the forecasted sales for the next year are as follows:

DRAFT

Quarter	Forecast (in thousand £)
Q1	200.97
Q2	256.07
Q3	220.49
Q4	183.63

□

Advantages of Method of Simple Averages

1. Compared to the method of monthly averages, this approach provides a more logical and statistically sound procedure for measuring seasonal variations, making it preferable in many contexts.

2. This method yields ratio-to-trend values for every available period, unlike the Ratio to Moving Average method, which may not provide indices for all periods due to its reliance on moving averages.

3. By focusing on trends rather than averages, this method captures underlying patterns more effectively, especially when analyzing long-term data.

Disadvantages of Method of Simple Averages

1. Because the Ratio to Trend method may not adapt as well to cyclical variations, the seasonal indices calculated using this method can be more biased compared to those obtained through the Ratio to Moving Average method.

2. The method can be more complex to implement compared to simpler methods, requiring a deeper understanding of statistical principles and data analysis techniques.

3. The effectiveness of this method hinges on the accurate estimation of trends. If the trend estimation is flawed, the resulting seasonal indices may also be inaccurate.

12.8.3 Ratio-to-Moving-Average Method

The Ratio-to-Moving-Average Method is a technique used to compute seasonal indices when both trend and seasonal effects are present in time series data. This method smooths out fluctuations by applying moving averages, and then calculates seasonal indices by comparing the actual data to the smoothed values.

DRAFT

Steps for the Ratio-to-Moving-Average Method

1. **Compute Moving Averages:** To remove the seasonal effect, calculate a centered moving average corresponding to the length of the seasonal cycle. For quarterly data, a 4-quarter moving average is used.

2. **Center the Moving Averages:** Center the moving averages to align them with the original data points by averaging adjacent moving averages.

3. **Calculate Ratios (Seasonal Factors):** For each quarter (or month), divide the actual observed value by the corresponding centered moving average and multiply by 100. This gives percentage ratios that reflect the seasonal effects.

4. **Average the Ratios:** For each quarter (or month), calculate the average of the ratios across all years. This average represents the seasonal index for that quarter or month.

5. **Adjust the Seasonal Indices:** Adjust the seasonal indices so their total sums to the appropriate value (e.g., 400 for quarterly data or 1200 for monthly data).

Problem 12.4. *Table 12.8 shows the quarterly sales for Toys International for the years 2012 through 2017. The sales are reported in millions of dollars. Determine a quarterly seasonal index using the ratio-to-moving-average method.*

Table 12.8: Toys International Quarterly Sales ($ million), 2012 to 2017

Year	Winter	Spring	Summer	Fall
2012	6.7	4.6	10.0	12.7
2013	6.5	4.6	9.8	13.6
2014	6.9	5.0	10.4	14.1
2015	7.0	5.5	10.8	15.0
2016	7.1	5.7	11.1	14.5
2017	8.0	6.2	11.4	14.9

DRAFT

Solution

There are six steps to determining the quarterly seasonal indexes.

Step 1: Refer to Table 12.9 for the following explanation. The first step involves calculating the four-quarter moving total for 2012. Beginning with the winter quarter of 2012, we sum the values: $6.7, $4.6, $10.0, and $12.7, resulting in a total of $34.0 million. The moving total is then "shifted" by including the

sales from spring, summer, and fall of 2012 along with the winter sales of 2013. This gives a new total of \$33.8 million, calculated as $4.6+10.0+12.7+6.5$. This process is repeated for the quarterly sales across the six years. The complete list of moving totals is displayed in column 2 of Table 12.9. Note that the total \$34.0 is placed between the spring and summer quarters of 2012, and the next total, \$33.8, is positioned between summer and fall 2012. Be sure to frequently check the calculations to avoid errors.

Step 2: Each moving total from column 2 is divided by 4 to compute the four-quarter moving average, which is listed in column 3. These averages are still placed between the quarters. For instance, the first average (8.500) is positioned between the spring and summer quarters of 2012.

Step 3: Next, the moving averages are centered. The first centered moving average is calculated by $(8.500 + 8.450)/2 = 8.475$ and is placed opposite summer 2012. Similarly, the second centered moving average is found by $(8.450 + 8.450)/2 = 8.450$. This method is used for all subsequent moving averages. As shown in column 4, each centered moving average corresponds to a specific quarter.

Step 4: The seasonal index for each quarter is determined by dividing the sales data in column 1 by the centered moving average from column 4. This index represents the ratio of the original sales data to the moving average. To elaborate, if we denote the time series as TCSI and the moving average as TC, dividing TCSI by TC gives us the seasonal component SI. For example, the seasonal index for the summer quarter of 2012 is 1.180, calculated as $10.0/8.475$.

Step 5: The specific seasonal indexes are organized in Table 12.10. This table will help us locate the specific seasonals for the corresponding quarters. The values 1.180, 1.130, 1.141, 1.126, and 1.143 all represent estimates of the typical seasonal index for the summer quarter. A reasonable method to find a typical seasonal index is to average these values in order to eliminate the irregular component. So we find the typical index for the summer quarter by

$$\frac{1.180 + 1.130 + 1.141 + 1.126 + 1.143}{5} = 1.144.$$

Step 6: The four quarterly means (0.767, 0.576, 1.144, and 1.522) should theoretically total 4.00 because the average is set at 1.0. The total of the four quarterly means may not exactly equal 4.00 due to rounding. In this problem, the total of the means is 4.009. A correction factor is therefore applied to each of the four means to force them to total 4.00.

Correction Factor for Adjusting Quarterly Means

DRAFT

Table 12.9: Computations Needed for the Specific Seasonal Indexes

Year	Quarter	(1) Sales	(2) Four-Quarter Total	(3) Four-Quarter Moving Average	(4) Centered Moving Average	(5) Specific Seasonal
2012	Winter	6.7	-	-	-	-
	Spring	4.6	34.0	8.500	-	-
	Summer	10.0	33.8	8.450	8.475	1.180
	Fall	12.7	33.8	8.450	8.450	1.503
2013	Winter	6.5	33.6	8.400	8.425	0.772
	Spring	4.6	34.5	8.625	8.513	0.540
	Summer	9.8	34.9	8.725	8.675	1.130
	Fall	13.6	35.3	8.825	8.775	1.550
2014	Winter	6.9	35.9	8.975	9.300	0.775
	Spring	5.0	36.4	9.100	9.038	0.553
	Summer	10.4	36.5	9.125	9.113	1.141
	Fall	14.1	37.0	9.250	9.188	1.535
2015	Winter	7.0	37.4	9.350	9.300	0.753
	Spring	5.5	38.3	9.575	9.463	0.581
	Summer	10.8	38.4	9.600	9.588	1.126
	Fall	15.0	38.6	9.650	9.625	1.558
2016	Winter	7.1	38.9	9.725	9.688	0.733
	Spring	5.7	38.4	9.600	9.663	0.590
	Summer	11.1	39.3	9.825	9.713	1.143
	Fall	14.5	39.8	9.950	9.888	1.466
2017	Winter	8.0	40.1	10.025	9.888	0.801
	Spring	6.2	40.5	10.125	10.075	0.615
	Summer	11.4	-	-	-	-
	Fall	14.9	-	-	-	-

$$\text{Correction Factor} = \frac{4.00}{\text{Total of four means}}$$

In this example,

$$\text{Correction factor} = \frac{4.00}{4.009} = 0.997755$$

The adjusted winter quarterly index is, therefore,

$$0.767 \times 0.997755 = 0.765.$$

Each of the seasonal means is adjusted downward so that their sum equals 4.00. Typically, indexes are presented as percentages, so each value in the final row

DRAFT

Table 12.10: Calculations Needed for Typical Quarterly Indexes

Year	**Winter**	**Spring**	**Summer**	**Fall**	**Total**
2012	-	-	1.180	1.503	
2013	0.772	0.540	1.130	1.550	
2014	0.775	0.553	1.141	1.535	
2015	0.753	0.581	1.126	1.558	
2016	0.733	0.590	1.143	1.466	
2017	0.801	0.615	-	-	
Total	3.834	2.879	5.720	7.612	
Mean	0.767	0.576	1.144	1.522	4.009
Adjusted	0.765	0.575	1.141	1.519	4.000
Index	76.5	57.5	114.1	151.9	

of Table 12.10 has been multiplied by 100.

For example, the index for the winter quarter is 76.5, while for the fall quarter, it is 151.9. These figures can be interpreted as follows: sales in the fall are 51.9% higher than the average quarter, whereas sales in the winter are 23.5% below the average quarter $(100.0 - 76.5)$. This pattern aligns with expectations—toy sales surge before Christmas (the fall quarter) and drop sharply after the holiday season (the winter quarter).

We now provide a concise explanation of the reasoning behind the previous calculations. The original data in the first column of Table 12.9 contain components for trend (T), cyclical (C), seasonal (S), and irregular (I) fluctuations. The main goal is to eliminate the seasonal component (S) from the original sales data.

Columns 2 and 3 of Table 12.9 are focused on calculating the centered moving averages shown in column 4. By averaging the values over four quarters, we smooth out the seasonal and irregular variations present in the original data from column 1. This results in values in column 4 that only reflect the trend and cyclical components (TC).

Subsequently, we divide the original sales data from column 1 (which contains TCSI) by the centered moving averages from column 4 (which contain TC). This division gives us the seasonal and irregular components (SI) in column 5. Mathematically, this is represented as TCSI/TC = SI. To express these

DRAFT

seasonal effects as indices, we multiply the resulting values by 100.

Advantages of Method of Ratio-to-Moving-Average

1. The seasonal indices calculated using this method exhibit less fluctuation compared to other methods, providing more stable estimates of seasonal patterns.

2. It allows for more precise seasonal adjustments by smoothing out irregularities and trends, making it particularly useful in time series analysis.

3. This method can be applied to different data frequencies (monthly, quarterly, etc.), enhancing its utility across various fields such as economics, finance, and inventory management.

4. By using moving averages, this method helps in identifying underlying trends alongside seasonal variations, which is beneficial for comprehensive data analysis.

Disadvantages of Method of Ratio-to-Moving-Average

1. Because it relies on moving averages, the results may lag behind current trends, which can be a disadvantage in rapidly changing environments where timely data is critical.

2. Calculating moving averages and ratios can be more complex than simpler methods, requiring more computational effort and understanding of statistical techniques.

3. Extreme values can disproportionately affect the moving averages, leading to biased seasonal indices if not properly managed.

4. The method may not perform well with short time series data, as there may not be enough observations to generate reliable moving averages.

12.8.4 Method of Link Relatives

The Link Relatives Method operates on the premise that trends are linear and cyclical variations exhibit a consistent pattern. Link relatives are calculated as percentages comparing the current period (whether a month or quarter) to the previous period. By calculating these link relatives and averaging them, the impact of cyclical and random components is reduced, and the trend is effectively removed during the adjustment of chain relatives.

The steps to compute seasonal indices using this method are as follows:

1. **Calculate Link Relatives (L.R.)**: For each period, divide the figure of that period by the figure from the previous period. For example, the

DRAFT

Link Relative for the 3rd quarter is computed as:

$$\text{L.R. of 3rd quarter} = \left(\frac{\text{Figure of 3rd quarter}}{\text{Figure of 2nd quarter}}\right) \times 100$$

2. **Average the Link Relatives**: Calculate the average of the link relatives for a specific quarter (or month) across different years. You can use either the arithmetic mean (A.M.) or the median (Md), with the median generally preferred to avoid skew from extreme values.

3. **Convert to Chained Relatives**: Assign a chained relative of 100 to the first quarter (or month). The chained relative (C.R.) for the current period is calculated as:

$$\text{C.R. for current period} = \left(\frac{\text{C.R. of previous period} \times \text{L.R. of current period}}{100}\right)$$

4. **Calculate C.R. for the First Quarter**: Determine the chained relative for the first quarter (or month) based on the last period's chained relative:

$$\text{C.R. of first quarter} = \left(\frac{\text{C.R. of last quarter} \times \text{average L.R. of first quarter}}{100}\right)$$

This value often differs from 100 due to long-term trends in the data.

5. **Adjust for Trend**: The chained relatives must be adjusted for the long-term trend effect. The adjustment factor (d) is calculated as:

- For quarterly data:

$$d = \frac{\text{C.R. for 1st quarter} - 100}{4}$$

- For monthly data:

$$d = \frac{\text{C.R. for 1st month} - 100}{12}$$

Assuming a linear trend, the adjustments $2d, 3d$, etc., are subtracted from the 2nd, 3rd, 4th quarters (or months) respectively.

6. **Calculate Seasonal Indices**: Express the adjusted chained relatives as a percentage of their average to derive seasonal indices.

7. **Check the Total**: Ensure that the total of these indices equals 400 for quarterly data and 1200 for monthly data.

The following problem and solution give the outlines the computation of seasonal indices using the Link Relatives method, demonstrated through an example dataset.

DRAFT

Problem 12.5. *Compute the seasonal indices by the 'Link Relatives' method for the adjoining data relating to the average quarterly prices (per kg) of a commodity for five years.*

Year	*Q1*	*Q2*	*Q3*	*Q4*
2020	*30*	*26*	*22*	*31*
2021	*35*	*28*	*22*	*36*
2022	*31*	*29*	*28*	*32*
2023	*31*	*31*	*25*	*35*
2024	*34*	*36*	*26*	*33*

Solution

Step 1: Calculate Link Relatives (L.R.)

The Link Relative for each quarter is calculated by comparing the value of each quarter to the corresponding quarter in the previous year.

$$\text{L.R.}_{Q1\,2020} = \frac{26}{30} \times 100 = 86.7$$
$$\text{L.R.}_{Q2\,2020} = \frac{22}{26} \times 100 = 84.6$$
$$\text{L.R.}_{Q3\,2021} = \frac{31}{22} \times 100 = 140.9$$
$$\text{L.R.}_{Q1\,2021} = \frac{35}{31} \times 100 = 112.9$$
$$\text{L.R.}_{Q2\,2021} = \frac{28}{35} \times 100 = 80.0$$
$$\text{L.R.}_{Q3\,2021} = \frac{22}{28} \times 100 = 78.6$$
$$\text{L.R.}_{Q4\,2021} = \frac{36}{22} \times 100 = 163$$

and so on.

All results are presented in Table 12.11.

Step 2: Calculate Totals for Link Relatives

Sum the values for each quarter:

DRAFT

Table 12.11: Computation of seasonal indices by link relatives method

Year	Q1	Q2	Q3	Q4	Average
2020	–	86.7	84.6	140.9	
2021	112.9	80.0	78.6	163.6	
2022	86.1	93.5	96.6	114.3	
2023	96.9	100.0	80.6	140.0	
2024	97.1	105.9	72.2	126.9	
Total	**393.0**	**379.4**	**328.0**	**544.8**	
Average LR	98.25	93.22	82.52	137.15	
Chain Relatives	**100**	93.22	76.93	105.50	
New CR	103.67				
Correction factor		0.92	1.83	2.75	
Adjusted CR	**100**	92.30	75.09	102.75	92.54
Seasonal Indices	108.06	99.75	81.15	111.04	

Total for Q1 = 393.0
Total for Q2 = 379.4
Total for Q3 = 328.0
Total for Q4 = 544.8

Step 3: Calculate Average Link Relatives

The average L.R. for each quarter is obtained by dividing the total by the number of years:

$$\text{Average L.R.}_{Q1} = \frac{393.0}{4} = 98.25$$
$$\text{Average L.R.}_{Q2} = \frac{379.4}{5} = 93.22$$
$$\text{Average L.R.}_{Q3} = \frac{328.0}{5} = 82.52$$
$$\text{Average L.R.}_{Q4} = \frac{544.8}{5} = 137.14$$

DRAFT

Step 4: Calculate Chain Relatives (C.R.)

Start with the first quarter set to 100 and compute the chain relatives for subsequent quarters:

$$\begin{aligned}
\text{C.R.}_{Q1} &= 100 \\
\text{C.R.}_{Q2} &= \frac{100 \times 93.22}{100} = 93.22 \\
\text{C.R.}_{Q3} &= \frac{93.22 \times 82.54}{100} \approx 76.93 \\
\text{C.R.}_{Q4} &= \frac{76.95 \times 137.14}{100} \approx 105.5
\end{aligned}$$

Step 5: Apply Correction Factor

To adjust for trends, calculate a correction factor based on the differences between the average L.R.s and C.R.s. The new (second) C.R for 1st quarter is:

$$\frac{\text{L.R of 1st Qrt} \times \text{C. R. of last (4th) Qrt}}{100} = \frac{105.5 \times 98.25}{100} = 103.67$$

Correction factor, $d = \frac{1}{4}(103.67 - 100) = 0.92$ Adjusted chain relatives are obtained by subtracting 0.92, 2×0.92=1.83, 3×0.92=2.75 from the chain relatives of second, third and fourth quarters respectively.

Step 6: Adjust the Chain Relatives

$$\begin{aligned}
\text{Average of Adjusted Chain Relatives} &= \frac{100 + 92.30 + 75.09 + 102.75}{4} \\
&= 92.54
\end{aligned}$$

The adjusted chain relatives are calculated using the correction factor:

$$\begin{aligned}
\text{Adjusted C.R.}_{Q1} &= 100 \\
\text{Adjusted C.R.}_{Q2} &\approx 92.30 \\
\text{Adjusted C.R.}_{Q3} &\approx 75.09 \\
\text{Adjusted C.R.}_{Q4} &\approx 102.75
\end{aligned}$$

Step 7: Calculate Seasonal Indices

Finally, the seasonal indices are expressed as percentages of their adjusted chain relatives:

$$\text{Seasonal Variation Index for any quarter} = \frac{\text{Adjusted C.R.}}{92.54} \times 100$$

DRAFT

$$\text{SI}_{Q1} = \frac{100}{98.25} \times 100 \approx 108.06$$
$$\text{SI}_{Q2} = \frac{92.30}{93.22} \times 100 \approx 99.75$$
$$\text{SI}_{Q3} = \frac{75.09}{82.52} \times 100 \approx 81.15$$
$$\text{SI}_{Q4} = \frac{102.75}{137.14} \times 100 \approx 111.04$$

Deseasonalization of Data

Deseasonalization is the process of removing seasonal effects from a time series to better analyze the underlying trends and patterns. This is particularly useful in data that exhibits regular fluctuations due to seasonal factors, such as sales data that may vary with seasons, holidays, or other recurring events.

Steps to Deseasonalize Data

1. **Collect Data:** Gather the time series data that contains seasonal patterns. Ensure it covers multiple cycles (e.g., several years of monthly data).

2. **Calculate Seasonal Indices:** Seasonal indices represent the relative level of each period compared to the average across all periods. To compute the seasonal indices, you can use any method explained above.

3. **Adjust Original Data:** Assuming a multiplicative model to describe the time series, the deseasonalized values are obtained by dividing the original values by their corresponding seasonal indices. That is,

$$\begin{aligned}\text{Deseasonalized Value} &= \frac{\text{Original Value}}{\text{Seasonal Index}} = \frac{y_t}{S_t} = \frac{T_t \times S_t \times C_t \times I_t}{S_t} \\ &= T_t \times C_t \times I_t\end{aligned}$$

Example of Deseasonalization

Consider the following Table 12.12 that contain the monthly sales data along with their corresponding seasonal indices:

Deseasonalize Data

The deseasonalized values (given in Table 12.13) are calculated by dividing the original sales figures by their corresponding seasonal indices:

DRAFT

Month	Sales	Seasonal Index
Jan	200	0.85
Feb	220	0.90
Mar	250	1.05
Apr	300	1.10
May	280	1.00
Jun	320	1.15
Jul	350	1.20
Aug	370	1.25
Sep	340	1.10
Oct	360	1.05
Nov	400	1.15
Dec	450	1.30

Table 12.12: Monthly Sales Data and Seasonal Indices

Month	Sales	Seasonal Index	Deseasonalized Sales
Jan	200	0.85	$\frac{200}{0.85} \approx 235.29$
Feb	220	0.90	$\frac{220}{0.90} \approx 244.44$
Mar	250	1.05	$\frac{250}{1.05} \approx 238.10$
Apr	300	1.10	$\frac{300}{1.10} \approx 272.73$
May	280	1.00	$\frac{280}{1.00} = 280.00$
Jun	320	1.15	$\frac{320}{1.15} \approx 278.26$
Jul	350	1.20	$\frac{350}{1.20} \approx 291.67$
Aug	370	1.25	$\frac{370}{1.25} = 296.00$
Sep	340	1.10	$\frac{340}{1.10} \approx 309.09$
Oct	360	1.05	$\frac{360}{1.05} \approx 342.86$
Nov	400	1.15	$\frac{400}{1.15} \approx 347.83$
Dec	450	1.30	$\frac{450}{1.30} \approx 346.15$

Table 12.13: Deseasonalized Sales Data

DRAFT

Importance of Deseasonalization

- **Trend Identification:** By removing seasonal effects, analysts can more easily identify long-term trends and cycles in the data.
- **Improved Forecasting:** Deseasonalized data can lead to more accurate forecasting models, as they are not influenced by predictable seasonal variations.
- **Meaningful Comparisons:** It allows for better comparisons between different time periods, making it easier to understand the true changes in data.

12.9 Measuring Cyclic Variation

Cyclical variation refers to fluctuations in a dataset that exhibit a tendency to rise and fall over time. After removing the trend and seasonal components from a time series, the remaining components are cyclical (C_t) and irregular (I_t). To effectively analyze the time series, it is crucial to measure and separate these cyclical fluctuations from the irregular variations. Business cycles, which represent significant fluctuations in economic data, typically span longer durations than seasonal variations.

Measuring cyclical variations is inherently challenging, as there is no universally accepted method for this purpose. The complexity arises because cyclical patterns are often intertwined with irregular factors that vary widely in timing, amplitude, and pattern. Consequently, deriving meaningful indices or fitting curves for cyclical variations is more difficult than for trend and seasonal components.

To address the irregular component I_t, one effective approach is to apply a moving average, usually calculated over a 3 or 5-month period. This technique, known as "smoothing," enhances the clarity of the cyclical variation within the data.

To effectively measure cyclic variation, several methods can be employed. The **residual method** is one of the popular methods, which is explained next.

12.9.1 Residual Method

The most commonly employed method for measuring cyclic variation is the **Residual Method**. This method is based on the multiple model of the time Series. This approach encompasses several systematic steps to isolate and analyze the cyclical component of a time series:

DRAFT

Steps for Measuring Cyclic Variation

When dealing with yearly data, there are no seasonal variations present. Therefore, the original data (y_t) are influenced by three main components:

- Trend (T_t)
- Cyclical (C_t)
- Irregular (I_t)

When monthly or quarterly data are given:

1. Estimate Trend and Seasonal Components: Begin by calculating the trend component T_t and seasonal values S_t for the time series data. The trend captures the long-term movement, while seasonal components account for regular fluctuations that repeat over time.

2. Calculate Components: Divide the original time series values Y_t by the estimated trend T_t and seasonal values S_t to obtain the cyclic component C_t and random component or irregular Movements I_t:

$$\frac{y_t}{T_t \times S_t} = \frac{T_t \times S_t \times C_t \times I_t}{T_t \times S_t} = C_t \times I_t.$$

3. Eliminate the Random Component: The random component I_t, similar to an error, cannot be removed directly. To reduce this component, we can use a moving average, usually calculated over a 3 or 5-month period. This technique, known as "smoothing," helps clarify the cyclical variation in the data.

DRAFT

Table 12.14: Computation of Cyclical Variation by using Residual Method.

Year and Quarter	Original Data (y_t)	Seasonal Index (S_t)	$\frac{y_t}{S_t} \times 100 = T_t \times C_t \times I_t$	Trend (T_t) (T_t)	$(C_t \times I_t)$ ×100	3-Quarter MA (C_t)
(1)	(2)	(3)	(4)	(5)	(6)=(4)/(5)	(7)
2021-Q1	274.4	108	254.1	242	105.0	–
2021-Q2	262.4	96	273.3	244	112.0	109.5
2021-Q3	224.9	82	274.3	246	111.5	108.0
2021-Q4	284.1	114	249.2	248	100.5	103.0
2022-Q1	261.9	108	242.5	250	97.0	98.0
2022-Q2	233.5	96	243.2	252	96.5	95.0
2022-Q3	190.6	82	232.4	254	91.5	94.0
2022-Q4	274.3	114	240.6	256	94.0	95.5
2023-Q1	281.4	108	260.6	258	101.0	97.5
2023-Q2	243.4	96	253.5	260	97.5	100.0
2023-Q3	218	82	265.9	262	101.5	103.5
2023-Q4	335.6	114	294.4	264	111.5	107.0
2024-Q1	310.3	108	287.3	266	108.0	110.0
2024-Q2	284.3	96	296.1	268	110.5	114.0
2024-Q3	273.5	82	333.5	270	123.5	107.3
2024-Q4	272.1	114	238.7	272	87.8	–

To illustrate the technique of time series decomposition, we start with the theoretical data presented in Table 12.14. Using the least square method, the trend equation for $y_t = 244.01 + 2.36t$ for each quarter $(t = 1, 2, \ldots, 16)$, we calculate the trend values, which are presented in Column (5). The data is first deseasonalized, then detrended, and finally smoothed using a three-period moving average to produce indices of cyclical variations. It is assumed that a seasonal variation index is available on a quarterly basis, and a least squares trend equation has already been calculated to provide the trend values by quarters, as shown in 12.14.

Remark 12.9.1. *The residual method is the most widely used approach for measuring cyclical variables. Other methods for analyzing cyclical variation include:*

(i). Reference Cycle Analysis Method

(ii). Direct Percentage Variation Method

(iii). Fitting of Sine Functions through Harmonic Analysis

12.10 Measuring Irregular Component

The irregular components of a time series indicate the residual fluctuations that persist after the trend, cyclical, and seasonal influences have been removed.

DRAFT

Therefore, when the original data is adjusted for T_t, S_t, and C_t, the resulting component is I_t. In practice, the cyclical variations are often so unpredictable and intertwined with irregular movements that effectively separating them can be extremely challenging.

12.11 Key Statistical Properties of Time Series

Understanding the statistical properties of time series data is crucial for accurate modeling and forecasting. Below are some key statistical properties.

12.11.1 Mean Function

The mean function of a time series $\{y_t; t = 1, 2, \ldots, n\}$ describes how the average value of the series changes over time. For a given time series $\{y_t\}$, the mean function is defined as the expected value of the series at time t:

$$\mu_t = E[y_t]$$

- If the time series is **stationary**, the mean function μ_t is constant, i.e., $\mu_t = \mu$ for all t.
- If the time series is **non-stationary**, the mean function μ_t can vary with time t, indicating trends or other changes in the average level of the series over time.

The mean function provides important insights into the underlying process generating the time series, especially when identifying trends or shifts in the average level over time.

12.11.2 Variance Function

The variance of a time series $\{y_t\}$ at time t is a measure of the dispersion or spread of the values around the mean. It is defined as the expected value of the squared deviation of y_t from its mean function μ_t:

$$\text{Var}(y_t) = E\left[(y_t - \mu_t)^2\right]$$

- If the time series is **stationary**, the variance $\text{Var}(y_t)$ is constant over time, i.e.,
$$\text{Var}(y_t) = \sigma^2 \quad \text{(a constant value for all } t\text{)}$$
- If the time series is **non-stationary**, the variance $\text{Var}(y_t)$ can change over time, indicating that the spread of the values around the mean varies with time. That is
$$\text{Var}(y_t) = \sigma_t^2$$

DRAFT

The variance function is crucial for understanding the stability of a time series, particularly in distinguishing between stationary and non-stationary processes.

12.11.3 Autocovariance Function

The autocovariance function of a time series $\{y_t\}$ measures the covariance between values of the series at different times t and $t + h$. It is defined as:

$$\gamma(h) = \text{Cov}(y_t, y_{t+h}) = E\left[(y_t - \mu_t)(y_{t+h} - \mu_{t+h})\right]$$

where:

- $\gamma(h)$ is the autocovariance function at lag h,
- y_t and y_{t+h} are values of the time series at times t and $t+h$ respectively,
- $\mu_t = E[y_t]$ and $\mu_{t+h} = E[y_{t+h}]$ are the mean functions of the time series at times t and $t + h$,

Interpretation

- If the time series is **stationary**, the autocovariance $\gamma(h)$ depends only on the lag h and not on the specific times t and $t + h$, so $\gamma(h)$ is the same for all t.
- If the time series is **non-stationary**, the autocovariance $\gamma(h)$ can depend on both the specific times t and $t + h$, as well as the lag h.

12.11.4 Autocorrelation Function

The autocorrelation function (ACF) of a time series $\{y_t\}$ measures the correlation between values of the series at different times t and $t + h$. It is defined as the normalized autocovariance function:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\text{Cov}(y_t, y_{t+h})}{\text{Var}(y_t)} = \frac{E\left[(y_t - \mu_t)(y_{t+h} - \mu_{t+h})\right]}{E\left[(y_t - \mu_t)^2\right]}$$

where:

- $\rho(h)$ is the autocorrelation function at lag h,
- $\gamma(h)$ is the autocovariance function at lag h,
- $\gamma(0)$ is the variance of the time series at time t,
- y_t and y_{t+h} are values of the time series at times t and $t+h$ respectively,
- $\mu_t = E[y_t]$ and $\mu_{t+h} = E[y_{t+h}]$ are the mean functions of the time series at times t and $t + h$.

DRAFT

Interpretation

- The autocorrelation $\rho(h)$ ranges from -1 to 1, where:
  - $\rho(h) = 1$ indicates perfect positive correlation,
  - $\rho(h) = -1$ indicates perfect negative correlation,
  - $\rho(h) = 0$ indicates no correlation.
- If the time series is **stationary**, the autocorrelation $\rho(h)$ depends only on the lag h and not on the specific times t and $t + h$.
- The autocorrelation function is crucial in identifying patterns within the time series, such as seasonality or trends.

The autocorrelation function provides insights into the internal structure of the time series by examining how past values influence future values over different time lags.

12.11.5 Correlogram

A **correlogram** is a graphical representation of the autocorrelation function of a time series. It plots the autocorrelation coefficients against different lags, providing a visual summary of the degree of correlation between the time series and its past values.

- **Lag:** The time difference between the observations being compared. For example, a lag of 1 compares each value with the one immediately preceding it.
- **Correlogram Plot:** On the plot, the x-axis represents the lags, and the y-axis represents the autocorrelation coefficients at those lags. See Figure 12.6. Confidence intervals are often displayed to help determine which autocorrelations are statistically significant.

Interpretation

- **Autocorrelation Values:** Values close to 1 or -1 indicate strong positive or negative autocorrelation, respectively.
- **Confidence Intervals:** If the autocorrelation at a particular lag falls outside the confidence interval, it is considered statistically significant.

The correlogram plots ρ_k against k, providing a visual representation of the autocorrelation at various lags.

DRAFT

Python Code for Drawing Correlogram

The correlogram of Gaussian white noise (details are given in Section 12.11.7), shown in Figure 12.6, reveals the following key findings:

- **Autocorrelation Values**: The autocorrelation values are close to zero across all lags, which is characteristic of white noise. This indicates that there are no significant correlations between the observations at different time points.
- **Significance Bounds**: Most autocorrelation values lie within the 95% confidence intervals (shadow area), suggesting that the observed correlations are statistically insignificant and likely due to random variation.
- **Randomness Confirmation**: The absence of significant spikes outside the confidence bounds supports the conclusion that the time series is purely random, with no underlying patterns or dependencies.

These findings confirm that the time series behaves as expected for a white noise process, where each observation is independent of the others.

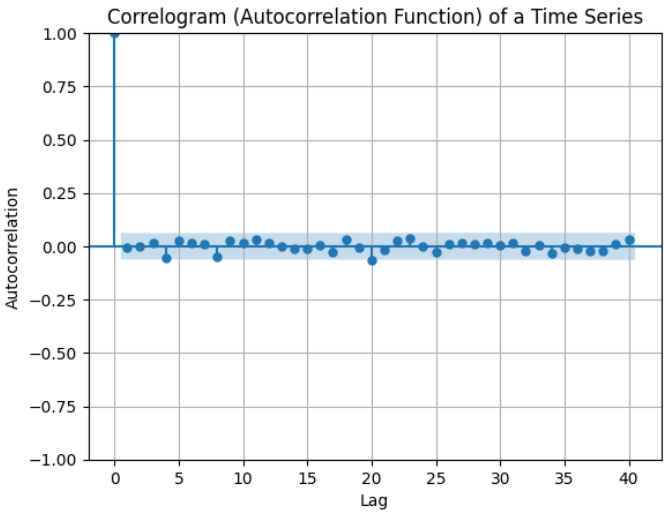


Figure 12.6: Correlogram of Normal Noise.

The following Python code generates a correlogram for a random time series, such as white noise:

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf

```

```
# Generate a random time series (e.g., white noise)
np.random.seed(42)
n = 1000
time_series = np.random.normal(0, 1, n)

# Plot the correlogram (ACF)
plt.figure(figsize=(10, 6))
plot_acf(time_series, lags=40, alpha=0.05)
plt.title("Correlogram (Autocorrelation Function) of a Time Series")
plt.xlabel("Lag")
plt.ylabel("Autocorrelation")
plt.grid(True)
plt.show()

```

Listing 12.7: Correlogram (Autocorrelation Function) of a Time Series

12.11.6 Partial Autocorrelation Function (PACF)

The **Partial Autocorrelation Function (PACF)** measures the correlation between a time series and its lagged values, after controlling for the effects of intermediate lags. For a stationary time series y_t, the partial autocorrelation at lag k is defined as:

$$\text{PACF}(k) = \text{corr}(y_t, y_{t-k} \mid y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})$$

The PACF is particularly useful in identifying the appropriate lag order of an autoregressive (AR) model. A significant partial autocorrelation at a particular lag indicates that the lag has a direct influence on the current observation, after accounting for the effects of other lags.

The **Partial Autocorrelation Function (PACF)** plot helps to understand the direct correlation between a time series and its lagged values, after accounting for the influence of intermediate lags. The PACF plot is particularly useful in identifying the appropriate lag order for autoregressive (AR) models.

Developing the formula for PACF

Let's rely on our $LAG = 2$ example for developing the PACF formula. Later, we'll generalize it to $LAG = k$. To know how much of the variance in $y_{(i-2)}$ has not been explained by the variance in $y_{(i-1)}$ we do two things:

- **Step 1:** We fit a linear regression model (i.e. a straight line) to the distribution of y_i versus $y_{(i-1)}$. This linear model will let us predict y_i from $y_{(i-1)}$. Conceptually, this linear model is allowing us to explain the variance in y_i as a function of the variance in $y_{(i-1)}$. But like all

DRAFT

optimally fitted models, our model is not going to be able to explain all of the variance in y_i. This fact takes us to step 2.

- **Step 2:** In this step, we calculate the residual errors of the linear model that we built in Step 1. The residual error is the difference between the observed value of y_i and the value predicted by the model. We do this residue calculation for each value of y_i so as to get a time series of residuals. This residuals time series gives us what we are looking for. It gives us the amount of variance in y_i which cannot be explained by the variance in $y_{(i-1)}$, plus of course some noise.

To calculate the second variable in the correlation, namely the amount of variance in $y_{(i-2)}$ that cannot be explained by the variance in $y_{(i-1)}$, we execute steps 1 and 2 above in the context of $y_{(i-2)}$ and $y_{(i-1)}$ instead of respectively y_i and $y_{(i-1)}$. This gives us the residuals series we are seeking for variable 2. So what we actually want to find out is the correlation between the following two variables:

- **Variable I:** The amount of variance in y_i that is not explained by the variance in $y_{(i-1)}$, and
- **Variable II:** The amount of variance in $y_{(i-2)}$ that is not explained by the variance in $y_{(i-1)}$.

This correlation is called the partial auto-correlation of y_i with $y_{(i-2)}$.

The final step is to apply the formula for Pearson's correlation coefficient to these two time series of residuals —the formula is

$$PACF(y_i, k=2) = \frac{Cov(y_i \mid y_{(i-1)}, y_{(i-2)} \mid y_{(i-1)})}{\sigma_{y_i|y_{(i-1)}} \times \sigma_{y_{(i-2)}|y_{(i-1)}}}$$

The general formula for $PACF(y, lag = k)$: In general, the partial auto-correlation of y_i with a k lagged version of itself i.e. $y_{(i-k)}$ is a correlation between the following two variables:

- **Variable 1:** the amount of variance in y_i that is not explained by the variance in $y_{(i-1)}, y_{(i-2)}, \ldots, y_{(i-k+1)}$, and,
- **Variable 2:** the amount of variance in $y_{(i-k)}$ that is not explained by the variance in $y_{(i-1)}, y_{(i-2)}, \ldots, y_{(i-k+1)}$.

In general formula

$$PACF(y_i, k) = \frac{Cov\begin{pmatrix}[y_i \mid y_{(i-1)}, y_{(i-2)}, \cdots, y_{(i-k+1)}], \\ [y_{(i-k)} \mid y_{(i-1)}, y_{(i-2)}, \cdots, y_{(i-k+1)}]\end{pmatrix}}{\sigma_{[y_i|y_{(i-1)}, y_{(i-2)}, \cdots, y_{(i-k+1)}]} \times \sigma_{[y_{(i-k)}|y_{(i-1)}, y_{(i-2)}, \cdots, y_{(i-k+1)}]}}$$

DRAFT

The use of this function was introduced as part of the **Box–Jenkins** approach to time series modelling, whereby plotting the partial autocorrelative functions one could determine the appropriate lags p in an $AR(p)$ model or in an extended $ARIMA(p, d, q)$ model.

Key Interpretation Points

- **Significant Lags**: Each bar in the PACF plot represents the partial autocorrelation at a specific lag. If a bar extends beyond the confidence interval, that lag is statistically significant. This indicates that the corresponding lag has a direct influence on the current observation, after controlling for the effects of shorter lags.
- **Cutoff Point**:
  - **Autoregressive (AR) Process**: In an AR process, the PACF typically shows significant correlations up to the lag order of the process, followed by a sharp cutoff. For example:
    - **AR(1) Process**: The PACF will show a significant spike at lag 1, with no significant correlations at higher lags.
    - **AR(2) Process**: The PACF will show significant spikes at lags 1 and 2, with no significant spikes at higher lags.
  - **Moving Average (MA) Process**: The PACF does not exhibit a sharp cutoff for an MA process.
  - **Mixed (ARMA) Process**: Neither the ACF nor the PACF will show a sharp cutoff.

Type of Process	Autocorrelation Function (ACF)	Partial Autocorrelation Function (PACF)
Autoregressive (AR) process	Tails off (decays gradually)	Cuts off sharply after lag p
Moving Average (MA) process	Cuts off sharply after lag q	Tails off (decays gradually)
Mixed (ARMA) process	Both its ACF and PACF will exhibit a tailing-off pattern.	Both its ACF and PACF will exhibit a tailing-off pattern.

- **Non-Stationarity**: If the PACF shows significant partial autocorrelations at many lags without a clear cutoff, this may indicate that the time series is non-stationary. In such cases, differencing the data may be necessary.

DRAFT

Example: Interpretation of a PACF Plot

Consider the following scenarios when interpreting a PACF plot:

- **AR(1) Process**: If the PACF plot shows a significant spike at lag 1, with no significant spikes at higher lags, this suggests that the time series follows an AR(1) process.
- **AR(2) Process**: If the PACF plot shows significant spikes at lags 1 and 2, with no significant correlations at higher lags, this indicates an AR(2) process.
- **No Clear Cutoff**: If the PACF does not show a sharp cutoff and has significant spikes at many lags, this could indicate non-stationarity or that an ARMA model might be more appropriate.

Example in Python

The following Python code computes and plots the PACF for a synthetic AR(1) process:

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_pacf

# Generate an example time series (AR(1) process)
np.random.seed(42)
n = 1000
phi = 0.5
y = np.zeros(n)
epsilon = np.random.normal(0, 1, n)

for t in range(1, n):
y[t] = phi * y[t-1] + epsilon[t]

# Plot the Partial Autocorrelation Function (PACF)
fig, ax = plt.subplots(figsize=(10, 6))
plot_pacf(y, ax=ax, lags=30)
plt.title("Partial Autocorrelation Function (PACF)")
plt.xlabel("Lag")
plt.ylabel("Partial Autocorrelation")
plt.grid(True)
plt.show()

```

Listing 12.8: Partial Autocorrelation Function (PACF) in Python

The PACF plot generated by this code will help determine the order of the autoregressive model. Significant spikes at specific lags suggest those lags should be included in the model.

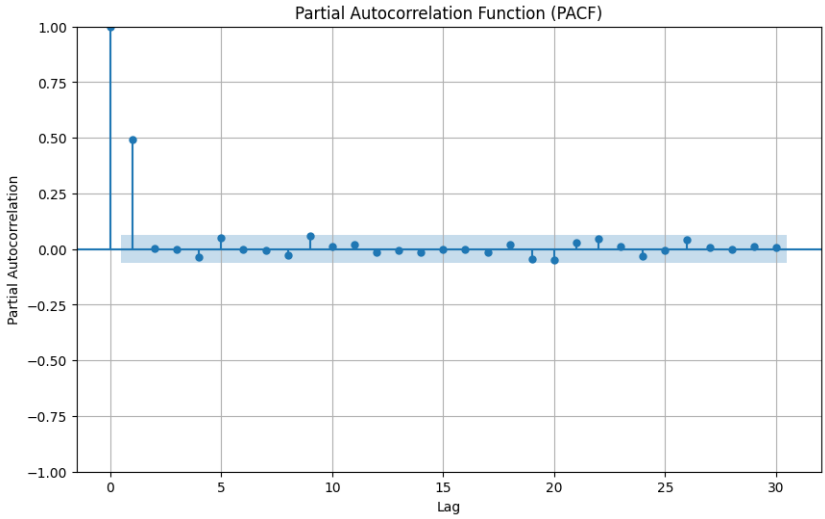


Figure 12.7: Partial Autocorrelation Function (PACF) by using simulated data.

12.11.7 White Noise

White noise is a type of **random signal** or process in which all possible values are **equally likely** and the signal exhibits **no predictable pattern**. In the context of time series analysis, white noise refers to a sequence of **uncorrelated** random variables with a **mean of zero** and a **constant finite variance**. A time series model ϵ_t which is weakly stationary with

$$E[\epsilon_t] = 0$$

$$\text{Cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

is said to be **white noise** with variance σ^2. Each observation in a white noise series is **independent** of all others, and there is no discernible relationship between the values at different points in time. Because of its randomness, white noise is often used as a benchmark in time series modeling to assess whether a time series model adequately captures the underlying data patterns. If the variables in the series are drawn from a Gaussian distribution, the series is called Gaussian white noise (also called normal white noise). For example, the plot of the Gaussian white noise with mean zero and variance one is depicted in the Figure 12.10.

DRAFT

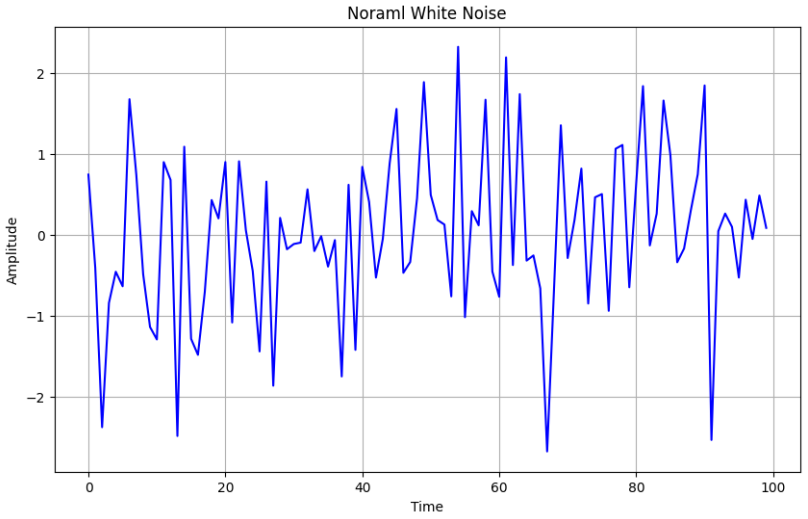


Figure 12.8: Normal White Noise with mean zero and variance one.

White noise is an important concept in time series forecasting. If a time series is white noise, it is a sequence of random numbers and cannot be forecasted. If the series of forecast errors are not white noise, it suggests improvements could be made to the predictive model.

```
# Normal White Noise
import numpy as np
import matplotlib.pyplot as plt

# Parameters for the white noise
n = 100  # Number of data points
mean = 0  # Mean of the white noise
std_dev = 1  # Standard deviation of the white noise

# Generate white noise
white_noise = np.random.normal(mean, std_dev, n)

# Plotting the white noise
plt.figure(figsize=(10, 6))
plt.plot(white_noise, color='blue')
plt.title("White Noise")
plt.xlabel("Time")
plt.ylabel("Amplitude")
plt.grid(True)
plt.show()


```

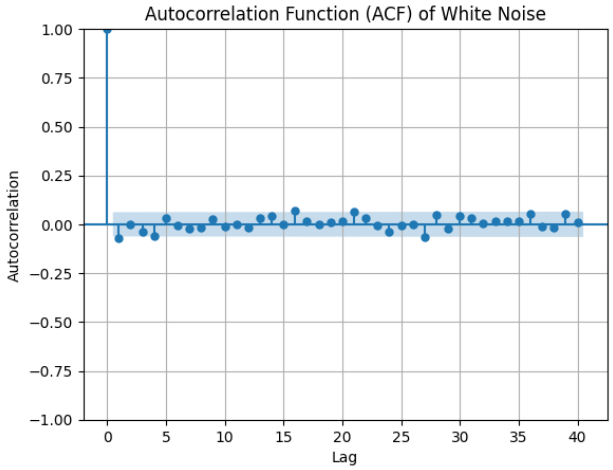


Figure 12.9: Autocorrelation fuention of Normal White Noise.

The autocorrelation function of normal white noise is presented in Figure 12.9.

```
# Plot the autocorrelation function (ACF)
plt.figure(figsize=(10, 4))
plot_acf(white_noise, lags=40, alpha=0.05)
plt.title("Autocorrelation Function (ACF) of White Noise")
plt.xlabel("Lag")
plt.ylabel("Autocorrelation")
plt.grid(True)
plt.show()


```

12.12 Stationary Processes

Stationary processes are a fundamental concept in time series analysis. A time series is considered **stationary** if its statistical properties, such as **mean**, **variance**, and **autocovariance**, are constant over time.

Example of a Stationary Time Series:

White noise is a classic example of a stationary time series. As mentioned earlier, white noise is a sequence of uncorrelated random variables with a constant mean of zero and a constant variance. Thus, the **mathematical model** of white noise time series is

$$y_t = \mu + \epsilon_t$$

where:

- μ is a constant (mean of the series).
- ϵ_t represents white noise, which has a mean of zero and constant variance σ^2.

For this time series:

- The **mean** is constant at $\mathbb{E}[y_t] = \mu$.
- The **variance** is constant at $\text{Var}(y_t) = \sigma^2$.
- The **autocovariance** between y_t and y_{t+k} is $\text{Cov}(y_t, y_{t+k}) = 0$ for all $k \neq 0$.

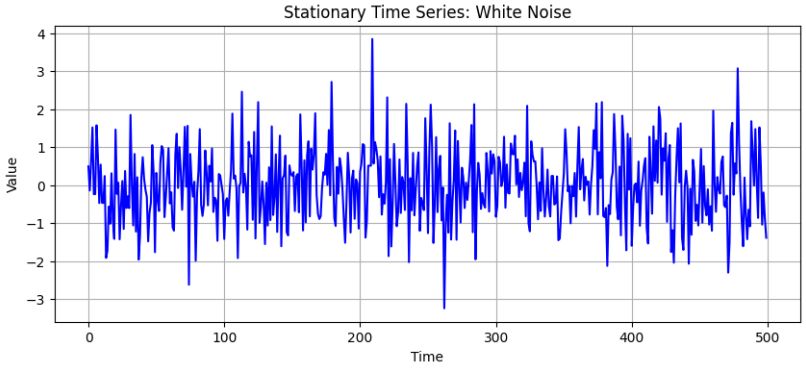


Figure 12.10: Stationary of time seris

Since the mean, variance, and autocovariance do not change over time, this time series is stationary. Figure 12.10 illustrates the stationary process of white noise. The plot shows random fluctuations around a constant mean (0 in this case), with no apparent trend or change in variance over time.

Types of stationarity

Stationarity can be defined in two ways: weak stationarity, which focuses solely on the covariance of a process, and strict stationarity, which is a more stringent condition that requires the distributions to remain consistent over time.

Definition 12.12.1 (Strictly Stationary)**.** *A time series* $\{y_t\}$ *is strictly stationary if the joint distribution of*

$$(y_{t_1}, y_{t_2}, \ldots, y_{t_k})$$

is the same as that of

$$(y_{t_1+h}, y_{t_2+h}, \ldots, y_{t_k+h})$$

for all time points $t_1, t_2, \ldots, t_k$ *and for all lags* h*. This implies that the entire distribution does not change over time.*

The assumption mentioned above is generally viewed as quite strong and can be difficult to verify with actual data. In many cases, it's possible to operate under a less stringent assumption known as weak or second-order stationarity.

Definition 12.12.2 (Weak (or Second-Order) Stationarity)**.** *A weaker form of stationarity, where the first two moments (mean and variance) are constant over time, and the autocovariance depends only on the lag between two time points and not on the actual time points themselves.*

12.13 Random Walk Model

A **random walk model** is a type of time series model where the value at any time step t is equal to the value at the previous time step plus a random shock or disturbance. This model is widely used in financial modeling, econometrics, and physics, as it represents a stochastic process that evolves over time in a manner that is unpredictable.

Mathematical Representation

A random walk model can be expressed as:

$$y_t = y_{t-1} + \epsilon_t$$

where:

- y_t is the value of the time series at time t.
- y_{t-1} is the value of the time series at time $t-1$.
- ϵ_t is a white noise error term (i.e., a random shock) at time t, typically assumed to be normally distributed with mean 0 and variance σ^2.

Properties:

- **Non-stationarity**: A random walk is generally non-stationary, meaning its statistical properties (like mean and variance) change over time.
- **No Mean Reversion**: Unlike some models that have a tendency to revert to a mean, a random walk does not revert to any particular value.
- **Cumulative Effect**: The impact of the random shocks accumulates over time, leading to a series that can drift far from its initial value.

DRAFT

Example in Python

The following Python code generates a random walk and plots it:

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
np.random.seed(42)
n = 1000  # Number of time steps
epsilon = np.random.normal(0, 1, n)  # White noise error terms

# Generate random walk
y = np.zeros(n)
for t in range(1, n):
y[t] = y[t-1] + epsilon[t]

# Plot the random walk
plt.figure(figsize=(10, 6))
plt.plot(y, label="Random Walk", color="blue")
plt.title("Random Walk Time Series")
plt.xlabel("Time")
plt.ylabel("Value")
plt.grid(True)
plt.legend()
plt.show()

```

Listing 12.9: Random Walk Simulation in Python

DRAFT

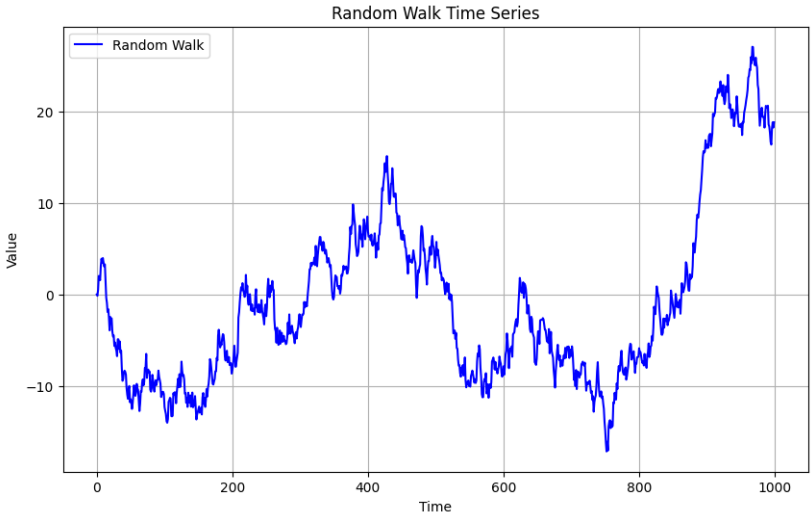


Figure 12.11: Random walk model by using simulated data.

Interpretation of the Plot

- **Unpredictability**: The path of the random walk shows that it can move in any direction at any time, depending on the random shocks.
- **Long-term Drift**: Over time, the random walk can drift away from the initial value due to the cumulative effect of the random shocks.

Applications

- **Financial Markets**: The random walk model is often used to model stock prices, which are assumed to follow a random walk in the Efficient Market Hypothesis.
- **Econometrics**: Random walks are used to model non-stationary economic variables, such as GDP or interest rates.

Random Walk Model with Drift

A **random walk model with drift** is an extension of the random walk model, where each new value is influenced by the previous value, a random shock, and a constant drift term. The drift term represents a systematic, deterministic trend that pushes the time series either upward or downward over time.

Mathematical Representation

The random walk with drift can be written as:

DRAFT

$$y_t = \mu + y_{t-1} + \epsilon_t$$

Where:

- y_t is the value of the time series at time t.
- μ is the drift term, representing a constant upward or downward trend.
- y_{t-1} is the value of the time series at time $t-1$.
- ϵ_t is the white noise error term at time t, usually assumed to be normally distributed with mean 0 and variance σ^2.

Characteristics

- **Non-Stationarity**: Like the simple random walk, the random walk with drift is typically non-stationary.
- **Trend**: The drift term introduces a linear trend into the time series. If $\mu > 0$, the series tends to increase over time. If $\mu < 0$, the series tends to decrease.

Example in Python

The following Python code simulates a random walk with drift and plots it:

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
np.random.seed(42)
n = 1000  # Number of time steps
mu = 0.1  # Drift term
epsilon = np.random.normal(0, 1, n)  # White noise error terms

# Generate random walk with drift
y = np.zeros(n)
for t in range(1, n):
y[t] = mu + y[t-1] + epsilon[t]

# Plot the random walk with drift
fig, ax = plt.subplots(figsize=(10, 6))
plt.plot(y, label="Random Walk with Drift", color="blue")
ax.set_xlabel("Time Steps")
ax.set_ylabel("Value")

# Add grid and legend
ax.grid(True)
```

DRAFT

```
ax.legend()

# Show the plot
plt.show()

```

Listing 12.10: Random Walk with Drift Simulation in Python



Figure 12.12: Random walk model with drift by using simulated data.

Interpretation

The drift term introduces a consistent upward or downward trend in the time series, which is overlaid on top of the random fluctuations caused by the white noise error terms. The process remains non-stationary because the mean of the series changes over time, driven by the drift.

12.14 Stationarity Test

When working with time series data, one of the fundamental concepts is stationarity. A time series is said to be stationary if its statistical properties, like mean and variance, do not change over time. Stationarity is crucial because many time series models assume that the data is stationary to make accurate forecasts and inferences.

Checking stationarity graphically involves visual inspection to identify patterns that might indicate whether a time series is stationary. Here we explain several tests that can be used to determine whether a time series is stationary:

- Dickey-Fuller Test
- Augmented Dickey-Fuller Test
- Phillips–Perron test
- Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test
- ADF-GLS test

12.14.1 Dickey-Fuller Test

The Dickey-Fuller test is used to test for the presence of a unit root in a time series. The test is based on the following regression model:

$$\Delta y_t = \alpha + \rho y_{t-1} + \epsilon_t$$

where:

- y_t is the time series at time t,
- $\Delta y_t = y_t - y_{t-1}$ is the first difference of the series,
- α is a constant (intercept term),
- ρ is the coefficient on the lagged level of the series,
- ϵ_t is the error term (assumed to be white noise).

The null hypothesis (H_0) is that the series has a unit root:

$$H_0 : \rho = 0$$

The alternative hypothesis (H_1) is that the series is stationary:

$$H_1 : \rho < 0$$

The test statistic is given by:

$$\text{ADF Statistic} = \frac{\widehat{\rho}}{\text{SE}(\widehat{\rho})}$$

where:

- $\widehat{\rho}$ is the estimated coefficient of y_{t-1} from the regression,
- $\text{SE}(\widehat{\rho})$ is the standard error of $\widehat{\rho}$.

DRAFT

To determine whether to reject the null hypothesis, compare the ADF statistic to critical values from the Dickey-Fuller distribution table. If the ADF statistic is less than (more negative than) the critical value, reject the null hypothesis of a unit root.

Note that the Dickey-Fuller (DF) test is designed to test for the presence of a unit root in a time series, and it does not directly use a t-test because the nature of the hypothesis being tested and the underlying assumptions are different from those of a standard t-test. Here's a detailed explanation:

- Dickey-Fuller Test: Assumes that the time series follows an autoregressive process and that the errors are not necessarily normally distributed. The test statistic is derived from the properties of the autoregressive model.
- t-Test: Assumes that the data are normally distributed (or the sample size is large enough for the Central Limit Theorem to apply) and that the errors are independently and identically distributed with constant variance.

12.14.2 Augmented Dickey-Fuller (ADF) Test

The **Augmented Dickey-Fuller (ADF) Test** is one of the most commonly used statistical tests to determine whether a time series is stationary. It tests for the presence of a unit root in the time series, which is an indicator of non-stationarity. The ADF test enhances the original Dickey-Fuller test by incorporating additional lagged differences of the time series to account for higher-order autocorrelation. This extension makes the ADF test more robust and flexible, enabling it to handle more complex time series dynamics and improve the reliability of the unit root test by addressing serial correlation that the simpler DF test might miss.

The ADF test is based on the following regression equation:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \epsilon_t \tag{12.4}$$

Where:

- Δy_t is the first difference of the time series, i.e., $\Delta y_t = y_t - y_{t-1}$.
- α is a constant term (intercept).
- βt is a trend term (optional).
- γ is the coefficient of the lagged level of the time series, y_{t-1}.
- δ_i are the coefficients of the lagged differences of the time series.

DRAFT

- ϵ_t is the white noise error term.
- p is the number of lags included to account for autocorrelation in the residuals.

Hypotheses

The hypotheses of the ADF test are as follows:

- **Null Hypothesis** (H_0): The time series has a unit root (i.e., the time series is non-stationary).
- **Alternative Hypothesis** (H_1): The time series does not have a unit root (i.e., the time series is stationary).

That is,

- Null Hypothesis (H_0): $\gamma = 0$ (unit root present),
- Alternative Hypothesis (H_1): $\gamma < 0$ (stationary).

Test Statistic

The test statistic for the ADF test is the t-statistic for the coefficient γ in the regression equation. Specifically, the test evaluates whether $\gamma = 0$, which would indicate the presence of a unit root.

Critical Values

The ADF test compares the test statistic to critical values at various significance levels (e.g., 1%, 5%, 10%). If the test statistic is less than the critical value, the null hypothesis is rejected, and the time series is considered stationary.

Decision Rule

- If the test statistic is less than the critical value, reject the null hypothesis. This implies the time series is stationary.
- If the test statistic is greater than the critical value, fail to reject the null hypothesis. This implies the time series is non-stationary.

Example with Python

We will use the Air Passengers dataset to perform the Dickey-Fuller test. This dataset contains the monthly totals of international airline passengers from 1949 to 1960.

DRAFT

```
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import adfuller

# Load Air Passengers data
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/airline-passengers.csv'
data = pd.read_csv(url, parse_dates=['Month'], index_col='Month')

# Display the first few rows of the dataset
print(data.head())

# Plot the time series
plt.figure(figsize=(12, 6))
plt.plot(data, label='Monthly Airline Passengers')
plt.title('Monthly International Airline Passengers')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()

# Perform the Dickey-Fuller test
result = adfuller(data['Passengers'])

print('ADF Statistic:', result[0])
print('p-value:', result[1])
print('Critical Values:', result[4])
```

Listing 12.11: Python Code for Dickey-Fuller Test

Output and Interpretation

The output of the Dickey-Fuller test for the Air Passengers dataset is as follows:

```
ADF Statistic: 0.8153688792060498
p-value: 0.991880243437641
Critical Values: {'1%': -3.482, '5%': -2.884, '10%': -2.578}
```

Interpretation

The ADF statistic is positive, indicating a failure to reject the null hypothesis of a unit root. The ADF statistic (0.815) is well above (less negative than) all critical values. This means the test statistic does not fall within the rejection region for the null hypothesis at any common significance level.

A high p-value (0.992) suggests that we fail to reject the null hypothesis of a unit root. This is much higher than common significance levels (e.g., 0.05

or 0.01). Based on the Dickey-Fuller test results, we cannot reject the null hypothesis of a unit root. This suggests that the time series is non-stationary. Further preprocessing, such as differencing, may be necessary to make the series stationary before applying models that assume stationarity.

12.14.3 Phillips-Perron (PP) Test

The Phillips-Perron (PP) test is another method for testing the presence of a unit root in a time series, similar to the ADF test. The PP test is designed to handle **heteroskedasticity** and **autocorrelation** in the errors of the time series data, making it a robust alternative to the ADF test. The PP test examines the following model:

$$y_t = \alpha + \beta t + \rho y_{t-1} + \epsilon_t$$

where,

- y_t is the time series,
- t is a time trend,
- α, β are parameters,
- ρ is the autoregressive parameter,
- ϵ_t is the error term.

Hypotheses:

- Null Hypothesis (H_0): $\rho = 1$ (unit root present),
- Alternative Hypothesis (H_1): $\rho < 1$ (stationary).

DRAFT

Example with Python

We will use the Air Passengers dataset to perform the PP test. This dataset contains the monthly totals of international airline passengers from 1949 to 1960.

```
pip install arch

import pandas as pd
import matplotlib.pyplot as plt
from arch.unitroot import PhillipsPerron

# Load Air Passengers data
url = 'https://raw.githubusercontent.com/jbrownlee/
  Datasets/master/airline-passengers.csv'
```

```
data = pd.read_csv(url, parse_dates=['Month'], index_col='Month')

# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(data.head())

# Plot the time series
plt.figure(figsize=(12, 6))
plt.plot(data, label='Monthly Airline Passengers')
plt.title('Monthly International Airline Passengers')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()

# Perform the Phillips-Perron test
pp_test = PhillipsPerron(data['Passengers'])

# Display the results of the PP test
print("\nPhillips-Perron Test Results:")
print('PP Statistic:', pp_test.stat)
print('p-value:', pp_test.pvalue)
print('Critical Values:', pp_test.critical_values)

# Interpret the results
print("\nInterpretation:")
if pp_test.pvalue < 0.05:
print("The p-value is less than 0.05. We reject the null hypothesis and conclude that the time series is stationary.")
else:
print("The p-value is greater than 0.05. We fail to reject the null hypothesis and conclude that the time series has a unit root (is non-stationary).")

```

Listing 12.12: Python Code for Phillips-Perron (PP) Test

The results of the Phillips-Perron test for the Air Passengers data are as follows:

- **PP Statistic**: -1.3511
- **p-value**: 0.6055
- **Critical Values**:
  - 1%: -3.4769
  - 5%: -2.8820

- 10%: -2.5777

Interpretation of Results:

- The PP statistic (-1.3511) is greater than the critical values at all significance levels (1%, 5%, 10%).
- The p-value (0.6055) is greater than the significance level of 0.05.

Since the PP statistic does not fall below any of the critical values and the p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, based on the Phillips-Perron test results, we conclude that the time series has a unit root and is non-stationary.

12.14.4 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used to test for the stationarity of a time series. Unlike the Dickey-Fuller tests, which test for the presence of a unit root (i.e., non-stationarity), the KPSS test evaluates the null hypothesis that a time series is stationary around a deterministic trend or level. The KPSS test examines the following model:

$$y_t = \alpha + \beta t + \eta_t$$

where:

- y_t is the time series,
- t is a time trend,
- α, β are parameters,
- η_t is a stationary error term.

Hypotheses:

- **Null Hypothesis (H_0)**: The time series is stationary around a deterministic trend (trend-stationary) or level (level-stationary).
- **Alternative Hypothesis (H_1)**: The time series has a unit root (i.e., it is non-stationary).

KPSS Test Procedure

The KPSS test involves the following steps:

1. Estimate the test statistic based on the residuals from a regression of the time series on a constant (for level stationarity) or a constant and a trend (for trend stationarity).
2. Compare the test statistic to critical values to determine whether to reject the null hypothesis.

DRAFT

Test Statistic

The KPSS test statistic is computed as follows:

$$\text{KPSS Statistic} = \frac{\sum_{t=1}^{T} \widehat{e}_t^2}{T \cdot \widehat{\sigma}^2}$$

where:

- $\widehat{e}_t$ are the residuals from the regression.
- $\widehat{\sigma}^2$ is the estimate of the long-run variance of the residuals.
- T is the number of observations.

Critical Values

The critical values for the KPSS test are determined based on the test statistic and depend on the chosen significance level (e.g., 1%, 5%, 10%).

Python Implementation and Results

Here is how you can perform the test in Python using the 'statsmodels' library: KPSS

```
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import kpss

# Load Air Passengers data
url = 'https://raw.githubusercontent.com/jbrownlee/
  Datasets/master/airline-passengers.csv'
data = pd.read_csv(url, parse_dates=['Month'], index_col='
  Month')

# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(data.head())

# Plot the time series
plt.figure(figsize=(12, 6))
plt.plot(data, label='Monthly Airline Passengers')
plt.title('Monthly International Airline Passengers')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()

# Perform the KPSS test
```

```
kpss_stat, p_value, lags, crit_values = kpss(data['Passengers'], regression='c')

# Display the results of the KPSS test
print("\nKPSS Test Results:")
print('KPSS Statistic:', kpss_stat)
print('p-value:', p_value)
print('Critical Values:', crit_values)

# Interpret the results
print("\nInterpretation:")
if p_value < 0.05:
print("The p-value is less than 0.05. We reject the null hypothesis and conclude that the time series is non-stationary.")
else:
print("The p-value is greater than 0.05. We fail to reject the null hypothesis and conclude that the time series is stationary.")
```

Listing 12.13: Python Code for KPSSTest

Interpretation of Results

After performing the KPSS test, you will obtain the following outputs:

- **KPSS Statistic**: 1.6513
- **p-value**: 0.01
- **Critical Values**: '10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739.

Decision: Since, the KPSS statistic is greater than the critical value at a given significance level, and the p-value is less than the chosen significance level (e.g., 0.05), reject the null hypothesis.

12.15 Models for Stationary Time Series

Models for stationary time series are used to analyze and forecast data that exhibit consistent statistical properties over time, such as constant mean and variance. Here's an overview of the key models commonly used for stationary time series:

12.15.1 Autoregressive (AR) Model of Order p

The Autoregressive (AR) model expresses the current value of the time series as a linear function of its past values. The AR(p) model can be formulated as:

DRAFT

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

where:

- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients.
- ϵ_t is white noise with zero mean and constant variance.

Properties:

- The model is stationary if the roots of the characteristic polynomial lie outside the unit circle.

Selecting the Order p in an AR(p) Model

The order p in an AR(p) model can be selected using various methods, including:

1. Partial Autocorrelation Function (PACF)

One of the most common methods to select p is by analyzing the Partial Autocorrelation Function (PACF). The PACF of an AR(p) process typically shows a significant spike at lag p and quickly diminishes to zero afterward. The steps are:

- Compute the PACF of the time series.
- Identify the lag p where the PACF cuts off (i.e., after which the PACF values are not significantly different from zero). Hence,
  - The PACF is non-zero for lags $k \leq p$.
  - The PACF drops to zero for lags $k > p$.
- The value of p where the PACF cuts off is typically selected as the order of the AR model.
- The PACF is non-zero for lags $k \leq p$.
- The PACF drops to zero for lags $k > p$.

Therefore, the order p of the AR model can be identified as the lag at which the PACF plot cuts off or drops significantly to near zero.

DRAFT

Example: Estimating p Using PACF

Let's illustrate the process of estimating p with an example. Assume we have a time series that follows an AR(2) process with coefficients $\phi_1 = 0.6$ and $\phi_2 = -0.2$. First, we simulate the AR(2) process using Python:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import pacf
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.tsa.arima_process import ArmaProcess

# Simulate an AR(2) process as an example
np.random.seed(42)
n = 500
phi1, phi2 = 0.6, -0.2
ar = np.array([1, -phi1, -phi2])  # AR coefficients (with lag 0 included)
ma = np.array([1])  # MA coefficients (just lag 0, as this is purely AR)

arma_process = ArmaProcess(ar, ma)
y = arma_process.generate_sample(nsample=n)

# Convert to pandas Series for convenience
y_series = pd.Series(y)

# Plot the PACF to visualize the partial autocorrelations
plot_pacf(y_series, lags=20, method='ywm')
plt.title('PACF Plot of Simulated AR(2) Process')
plt.show()

# Compute the PACF values
pacf_values = pacf(y_series, nlags=20)

# Print PACF values for first few lags
print("PACF values up to lag 20:")
for lag, value in enumerate(pacf_values):
print(f"Lag {lag}: {value:.4f}")

# The lag where PACF drops off indicates the order p of the AR model.
```

The PACF plot for the simulated AR(2) process is shown in Figure 12.13. In this plot, you can observe the following:

- Significant spikes at lags 1 and 2.
- The PACF drops to near zero after lag 2.

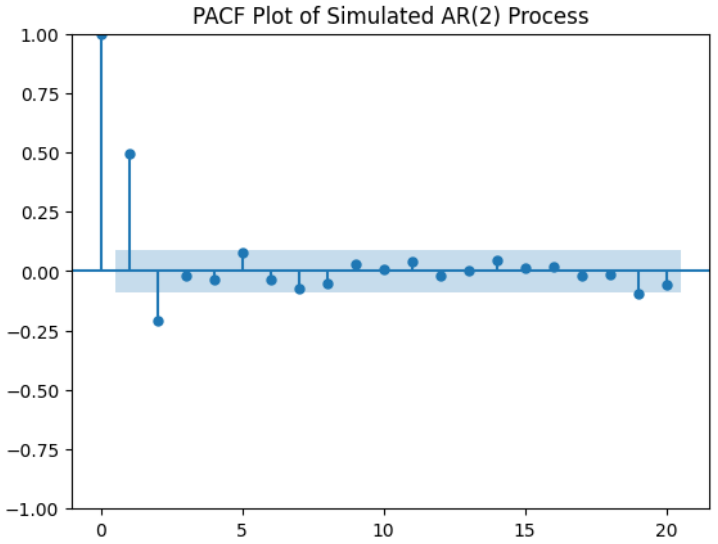


Figure 12.13: PACF Plot of Simulated AR(2) Process

Based on this behavior, we conclude that the appropriate order p for the AR model is 2.

2. Information Criteria

Information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are commonly used to select the optimal order p:

- **Akaike Information Criterion (AIC)**:

$$\text{AIC}(p) = -2\ln(\widehat{L}) + 2p,$$

  where $\widehat{L}$ is the maximum likelihood of the model.

- **Bayesian Information Criterion (BIC)**:

$$\text{BIC}(p) = -2\ln(\widehat{L}) + p\ln(n),$$

  where n is the number of observations.

- Choose the order p that minimizes the AIC or BIC.

The choice between AIC and BIC depends on the specific goals and context of your analysis. Both criteria are used for model selection, but they have different strengths and trade-offs. If interpretability and parsimony are critical, BIC might be better. If prediction is key, and you have enough data to support a more complex model, AIC could be the better option.

DRAFT

Example: Estimating p Using BIC

Again we follow an AR(2) process with coefficients $\phi_1 = 0.6$ and $\phi_2 = -0.2$.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.ar_model import AutoReg
from statsmodels.tsa.arima_process import ArmaProcess

# Simulate an AR(2) process as an example
np.random.seed(42)
n = 500
phi1, phi2 = 0.6, -0.2
ar = np.array([1, -phi1, -phi2])  # AR coefficients (with
    lag 0 included)
ma = np.array([1])  # MA coefficients (just lag 0, as this
    is purely AR)

arma_process = ArmaProcess(ar, ma)
y = arma_process.generate_sample(nsample=n)

# Convert to pandas Series for convenience
y_series = pd.Series(y)

# Define the maximum p to consider
max_p = 10

# Initialize lists to store AIC and BIC values
aic_values = []
bic_values = []

# Fit AR models for different values of p and calculate AIC
    and BIC
for p in range(1, max_p + 1):
model = AutoReg(y_series, lags=p).fit()
aic_values.append(model.aic)
bic_values.append(model.bic)

# Plot AIC and BIC values
plt.figure(figsize=(10, 6))
plt.plot(range(1, max_p + 1), aic_values, marker='o', label=
    'AIC')
plt.plot(range(1, max_p + 1), bic_values, marker='o', label=
    'BIC')
plt.xlabel('Order p')
plt.ylabel('Information Criteria')
plt.title('AIC and BIC for AR(p) Model Selection')
plt.legend()
plt.show()
```

```
42
43 # Select the p with the minimum AIC and BIC
44 best_p_aic = np.argmin(aic_values) + 1
45 best_p_bic = np.argmin(bic_values) + 1
46
47 print(f"Optimal p based on AIC: {best_p_aic}")
48 print(f"Optimal p based on BIC: {best_p_bic}")
```

From Figure 12.14, we see the optimal p based on AIC is 10. On the other hand the optimal p based on BIC is 2. BIC tends to favor simpler models with fewer parameters, making it more conservative.

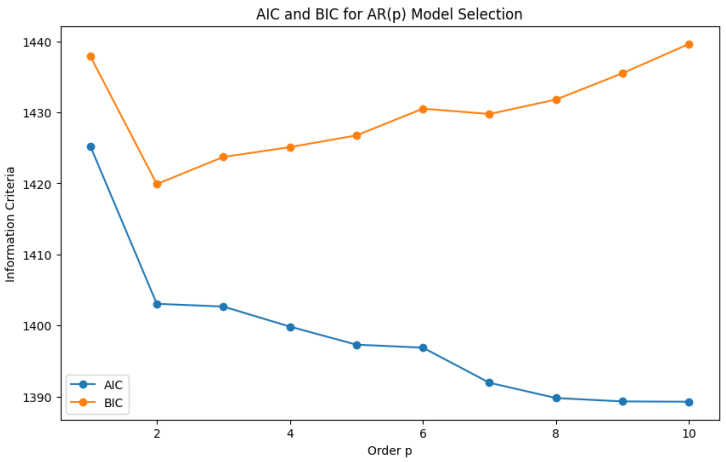


Figure 12.14: Information Criteria vs p Plot of Simulated AR(2) Process

3. Likelihood Ratio Test (LRT)

The Likelihood Ratio Test (LRT) can be used to compare nested models (i.e., AR(p) versus AR(p+1)):

- Fit AR(p) and AR(p+1) models.
- Calculate the likelihood ratio statistic:

$$\lambda = 2\left(\ln(\widehat{L}_{p+1}) - \ln(\widehat{L}_p)\right),$$

  where $\widehat{L}_{p+1}$ and $\widehat{L}_p$ are the maximum likelihoods of the AR(p+1) and AR(p) models, respectively.
- Compare λ with the critical value from the chi-square distribution with 1 degree of freedom.
- If λ is significant, choose the higher order; otherwise, choose p.

4. Cross-Validation

Cross-validation can be used to assess the predictive accuracy of models with different values of p:

- Split the data into training and test sets.
- Fit AR(p) models for a range of values of p on the training set.
- Compute the prediction error on the test set.
- Select the order p that minimizes the prediction error.

5. Residual Analysis

After fitting an AR(p) model, analyze the residuals:

- Fit models with different values of p.
- Analyze the residuals to check if they resemble white noise (i.e., no significant autocorrelation).
- Choose p such that the residuals show minimal autocorrelation and resemble white noise.

12.15.2 Example: AR(1) Model

The AR(1) model, or first-order autoregressive model, is defined as:

$$y_t = \phi y_{t-1} + \epsilon_t$$

where:

- y_t is the value of the time series at time t,
- ϕ is the autoregressive parameter,
- y_{t-1} is the value of the time series at time $t-1$,
- ϵ_t is the white noise error term with mean 0 and variance σ^2.

Stationarity

The AR(1) model is stationary if $|\phi| < 1$.

Mean and Variance

For a stationary AR(1) process, the mean is $E[y_t] = 0$ and the variance is given by:

$$\text{Var}(y_t) = \frac{\sigma^2}{1 - \phi^2}$$

DRAFT

Autocorrelation Function (ACF)

The autocorrelation at lag h is:

$$\rho(h) = \phi^h$$

Theorem 12.1. *For a stationary AR(1) process defined by*

$$y_t = \phi y_{t-1} + \epsilon_t,$$

where ϵ_t is white noise with mean 0 and variance σ^2, the variance of y_t is given by:

$$Var(y_t) = \frac{\sigma^2}{1 - \phi^2}.$$

Proof:

Step 1: Define the variance of y_t.

The variance of y_t is defined as:

$$\text{Var}(y_t) = E[y_t^2] - (E[y_t])^2.$$

Since the AR(1) process is stationary and has a mean of 0 (i.e., $E[y_t] = 0$), the variance simplifies to:

$$\text{Var}(y_t) = E[y_t^2].$$

Step 2: Substitute the AR(1) model equation into the variance expression.

Substitute the AR(1) equation $y_t = \phi y_{t-1} + \epsilon_t$ into the expression for $E[y_t^2]$:

$$E[y_t^2] = E[(\phi y_{t-1} + \epsilon_t)^2].$$

Step 3: Expand the expression.

Expand the square:

$$E[y_t^2] = E[\phi^2 y_{t-1}^2 + 2\phi y_{t-1}\epsilon_t + \epsilon_t^2].$$

Step 4: Simplify using properties of expectations.

Because y_{t-1} and ϵ_t are independent, the cross-term $E[2\phi y_{t-1}\epsilon_t]$ is 0 (since $E[\epsilon_t] = 0$). This simplifies the expression to:

$$E[y_t^2] = \phi^2 E[y_{t-1}^2] + E[\epsilon_t^2].$$

Given that $E[\epsilon_t^2] = \sigma^2$, we have:

$$E[y_t^2] = \phi^2 E[y_{t-1}^2] + \sigma^2.$$

Step 5: Use stationarity to solve for $E[y_t^2]$.

DRAFT

Since the process is stationary, $\text{Var}(y_t) = \text{Var}(y_{t-1}) = E[y_t^2]$. Let $\text{Var}(y_t) = \gamma(0)$:

$$\gamma(0) = \phi^2\gamma(0) + \sigma^2.$$

Step 6: Solve for $\gamma(0)$.
Rearranging the equation to solve for $\gamma(0)$:

$$\gamma(0)(1 - \phi^2) = \sigma^2.$$

Thus, the variance of y_t is:

$$\text{Var}(y_t) = \gamma(0) = \frac{\sigma^2}{1 - \phi^2}.$$

□

Theorem 12.2. *For a stationary AR(1) process defined by*

$$y_t = \phi y_{t-1} + \epsilon_t,$$

where ϵ_t is white noise with mean 0 and variance σ^2, the autocorrelation function (ACF) at lag h is given by:

$$\rho(h) = \phi^h \quad \text{for } h = 0, 1, 2, \ldots$$

Proof:

To prove this, we begin by calculating the autocovariance function $\gamma(h)$ of the AR(1) process.

Step 1: Express $\gamma(h)$ using the definition of covariance.
The covariance between y_t and y_{t+h} is:

$$\gamma(h) = \text{Cov}(y_t, y_{t+h}) = E[(y_t - \mu)(y_{t+h} - \mu)],$$

where $\mu = E[y_t]$ is the mean of the process. For a stationary process, $\mu = 0$ without loss of generality, so:

$$\gamma(h) = E[y_t y_{t+h}].$$

Step 2: Use the AR(1) model to express y_{t+h} in terms of y_t and the noise terms.
From the AR(1) model:

$$y_{t+1} = \phi y_t + \epsilon_{t+1}.$$

Applying this recursively, we can express y_{t+h} as:

$$y_{t+h} = \phi^h y_t + \sum_{i=1}^{h} \phi^{h-i}\epsilon_{t+i}.$$

DRAFT

Step 3: Calculate $\gamma(h)$ using the expectation.
Substituting the expression for y_{t+h} into the autocovariance function:

$$\gamma(h) = E\left[y_t\left(\phi^h y_t + \sum_{i=1}^{h} \phi^{h-i}\epsilon_{t+i}\right)\right].$$

Expanding this:

$$\gamma(h) = \phi^h E[y_t^2] + E\left[y_t \sum_{i=1}^{h} \phi^{h-i}\epsilon_{t+i}\right].$$

Since y_t is independent of ϵ_{t+i} (for $i \geq 1$), the second term equals 0:

$$\gamma(h) = \phi^h E[y_t^2] = \phi^h \gamma(0).$$

Step 4: Normalize to obtain the autocorrelation function.
The autocorrelation function $\rho(h)$ is obtained by normalizing $\gamma(h)$ with $\gamma(0)$:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\phi^h \gamma(0)}{\gamma(0)} = \phi^h.$$

Thus, the autocorrelation function of the AR(1) process is:

$$\rho(h) = \phi^h \quad \text{for } h = 0, 1, 2, \ldots$$

□

12.15.3 Parameter Estimation of AR(p) Model

1. Method of Least Squares (Ordinary Least Squares, OLS)

In the OLS method, the AR(p) model can be written as a linear regression equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\phi} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y} = (y_p, y_{p+1}, \ldots, y_T)$ is the vector of observations (excluding the first p values),
- $\mathbf{X}$ is the design matrix containing the lagged values of the series

$$(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$$

,

- $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)$ is the vector of autoregressive parameters to be estimated,

DRAFT

- $\boldsymbol{\epsilon}$ is the error term.

The OLS estimator for $\boldsymbol{\phi}$ is given by:

$$\widehat{\boldsymbol{\phi}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This method minimizes the sum of squared residuals $\sum_{t=p+1}^{T} \epsilon_t^2$ to find the estimates of the parameters.

2. Yule-Walker Equations

The Yule-Walker method uses autocovariances to estimate the parameters. It exploits the fact that the autocovariances $\gamma_k = \text{Cov}(y_t, y_{t+k})$ of an AR(p) process can be related to the AR coefficients.

For an AR(p) process, the Yule-Walker equations are:

$$\gamma_k = \sum_{i=1}^{p} \phi_i \gamma_{k-i}, \quad k = 1, 2, \ldots, p$$

Where:

- γ_k is the autocovariance at lag k.

To estimate $\phi_1, \ldots, \phi_p$, the Yule-Walker equations can be written in matrix form as:

$$\boldsymbol{\Gamma}\boldsymbol{\phi} = \boldsymbol{\gamma}$$

Where:

- $\boldsymbol{\Gamma}$ is the $p \times p$ autocovariance matrix, with entries $\gamma_{|i-j|}$,
- $\boldsymbol{\gamma}$ is the vector of autocovariances at lags $1, 2, \ldots, p$.

The solution to this system of linear equations gives the estimates of $\phi_1, \phi_2, \ldots, \phi_p$.

3. Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a general approach that maximizes the likelihood of the observed data given the model. Assuming the error terms ϵ_t are i.i.d. Gaussian, the log-likelihood function for the AR(p) model is:

$$L(\boldsymbol{\phi}, \sigma^2) = \prod_{t=p+1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon_t^2}{2\sigma^2}\right)$$

Where:

- σ^2 is the variance of the error term ϵ_t,
- $\epsilon_t = y_t - (\phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p})$.

DRAFT

The log-likelihood function is:

$$\ell(\boldsymbol{\phi}, \sigma^2) = -\frac{T-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}\epsilon_t^2$$

The parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)$ and σ^2 are estimated by maximizing this log-likelihood function.

4. Model Selection (Choosing p)

The order p of the AR model is typically chosen using information criteria like:

Akaike Information Criterion (AIC)

$$\text{AIC} = \ln(\widehat{\sigma}^2) + \frac{2p}{T}$$

Bayesian Information Criterion (BIC)

$$\text{BIC} = \ln(\widehat{\sigma}^2) + \frac{p\ln(T)}{T}$$

These criteria penalize models with more parameters to avoid overfitting.

Alternatively, the **Partial Autocorrelation Function (PACF)** can be used to choose p, as the PACF of an AR(p) process cuts off after lag p.

Example of Parameter Estimation Using Yule-Walker

For an AR(2) model, the Yule-Walker equations can be written as:

$$\begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

Solving this system gives the estimates $\widehat{\phi}_1$ and $\widehat{\phi}_2$.

12.15.4 Moving Average (MA) Model

The Moving Average (MA) model represents the current value of the series as a function of past error terms. The MA(q) model is given by:

$$y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

where:

- $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients.
- μ is the mean of the series.

DRAFT

- ϵ_t is white noise.

Properties:

- The model is stationary if the number of terms q is finite and the roots of the polynomial are outside the unit circle.

Estimating the Order q of an MA(q) Model

The order q of a Moving Average (MA) model can be estimated using various methods, such as analyzing the Autocorrelation Function (ACF) or utilizing Information Criteria like AIC and BIC. In this document, we demonstrate how to estimate q for an MA(q) model using both approaches and provide an interpretation of the results.

Autocorrelation Function (ACF)

For an MA(q) model, the autocorrelation function (ACF) exhibits non-zero values up to lag q, after which it drops to zero. This characteristic can be used to estimate the order q. Below is a sample ACF plot generated from a simulated MA(2) process:

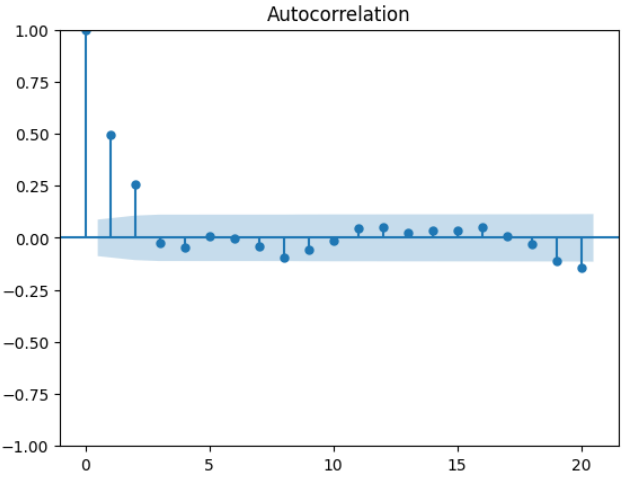


Figure 12.15: ACF plot of a simulated MA(2) process.

Interpretation: In the ACF plot shown in Figure 12.17, we observe significant spikes at lags 1 and 2. These spikes correspond to the non-zero autocorrelations of the MA(2) process. Beyond lag 2, the autocorrelation values drop to near zero, indicating that the order of the MA model is likely $q = 2$.

DRAFT

Information Criteria (AIC and BIC)

To further validate the order q, we can use Information Criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). These criteria balance the model fit with model complexity. The order q that minimizes AIC or BIC is considered the best estimate.

The AIC and BIC values are calculated for various orders q and plotted as shown below:

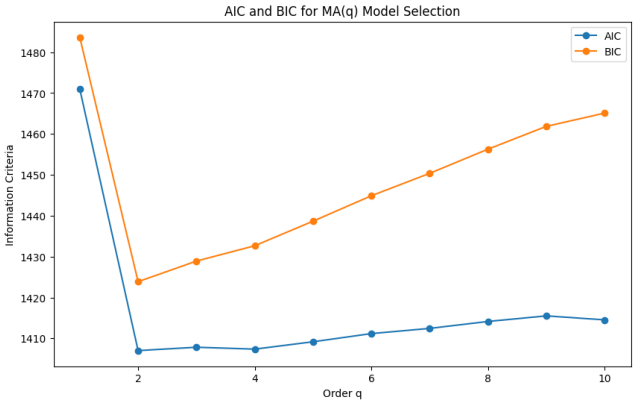


Figure 12.16: AIC and BIC values for different orders q of an MA(q) model.

Interpretation: In Figure 12.19, the AIC and BIC values are plotted for different orders q. We observe that both AIC and BIC reach their minimum values at $q = 2$, indicating that an MA(2) model is the best fit for the data.

Python Code

The following Python code was used to generate the ACF plot and calculate the AIC and BIC values:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_process import ArmaProcess
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.tsa.arima.model import ARIMA

# Simulate an MA(2) process as an example
np.random.seed(42)
theta1, theta2 = 0.5, 0.4
ar = np.array([1])  # AR part, since this is purely MA
ma = np.array([1, theta1, theta2])  # MA coefficients (
  with lag 0 included)

```

DRAFT

```
arma_process = ArmaProcess(ar, ma)
y = arma_process.generate_sample(nsample=500)

# Convert to pandas Series for convenience
y_series = pd.Series(y)

# Plot the Autocorrelation Function (ACF)
plot_acf(y_series, lags=20)
plt.show()

# Estimate q using Information Criteria (AIC and BIC)
max_q = 10
aic_values = []
bic_values = []

for q in range(1, max_q + 1):
model = ARIMA(y_series, order=(0, 0, q)).fit()
aic_values.append(model.aic)
bic_values.append(model.bic)

# Plot AIC and BIC values
plt.figure(figsize=(10, 6))
plt.plot(range(1, max_q + 1), aic_values, marker='o',
  label='AIC')
plt.plot(range(1, max_q + 1), bic_values, marker='o',
  label='BIC')
plt.xlabel('Order q')
plt.ylabel('Information Criteria')
plt.title('AIC and BIC for MA(q) Model Selection')
plt.legend()
plt.show()

# Select the q with the minimum AIC and BIC
best_q_aic = np.argmin(aic_values) + 1
best_q_bic = np.argmin(bic_values) + 1

print(f"Optimal q based on AIC: {best_q_aic}")
print(f"Optimal q based on BIC: {best_q_bic}")
```

Listing 12.14: Python Code for Estimating q of MA(q) Model

12.15.5 Example: MA(1) Model

The MA(1) model is defined as:

$$y_t = \epsilon_t + \theta \epsilon_{t-1},$$

where:

- y_t is the value of the time series at time t,

- ϵ_t is white noise with mean 0 and variance σ^2,
- θ is the parameter of the MA(1) model.

Moving Variance of MA(1) Model

The variance of y_t is:

$$\text{Var}(y_t) = \text{Var}(\epsilon_t + \theta\epsilon_{t-1}).$$

Expanding the variance:

$$\text{Var}(y_t) = \text{Var}(\epsilon_t) + \theta^2\text{Var}(\epsilon_{t-1}) + 2\theta\text{Cov}(\epsilon_t, \epsilon_{t-1}).$$

Since ϵ_t is white noise, $\text{Var}(\epsilon_t) = \sigma^2$, $\text{Var}(\epsilon_{t-1}) = \sigma^2$, and $\text{Cov}(\epsilon_t, \epsilon_{t-1}) = 0$, we get:

$$\text{Var}(y_t) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1 + \theta^2).$$

Autocorrelation Function (ACF) of MA(1) Model

The autocovariance function $\gamma(h)$ at lag h is:

$$\gamma(0) = \text{Var}(y_t) = \sigma^2(1 + \theta^2),$$

$$\gamma(1) = \theta\sigma^2,$$

$$\gamma(h) = 0 \quad \text{for } h > 1.$$

The autocorrelation function $\rho(h)$ is given by:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)},$$

so:

$$\rho(0) = 1, \quad \rho(1) = \frac{\theta}{1 + \theta^2}, \quad \rho(h) = 0 \text{ for } h > 1.$$

12.15.6 Autoregressive Moving Average (ARMA) Model

The ARMA(p, q) model combines both autoregressive (AR) and moving average (MA) components and is used to describe a stationary time series. It is defined as:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

where:

DRAFT

- y_t is the value of the time series at time t,
- μ refers to the mean of the original (non-differenced) process,
- $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the autoregressive (AR) part, representing the relationship between y_t and its past values $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$,
- $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the moving average (MA) part, representing the relationship between y_t and past error terms $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$,
- ϵ_t is a white noise error term at time t, assumed to be independently and identically distributed with mean zero and constant variance,
- p is the order of the autoregressive part, i.e., the number of lagged values of y_t included in the model,
- q is the order of the moving average part, i.e., the number of lagged error terms included in the model.

The model assumes that the time series y_t is **stationary**, meaning its statistical properties (such as mean, variance, and autocorrelation) do not change over time. The error terms ϵ_t are assumed to be **white noise**, meaning they are uncorrelated and have a constant variance. The ARMA model is widely used in time series forecasting and analysis, particularly for stationary data.

Properties of ARMA Model

The ARMA model has several important properties, which are summarized below:

- **Stationarity:** For an ARMA(p, q) model to be valid, the underlying time series y_t must be **stationary**. This means that the mean, variance, and autocovariance of the series do not depend on time.
- **Invertibility:** The MA part of the model requires the process to be **invertible**, meaning that the model's moving average representation can be written as an infinite autoregressive process. This is essential for forecasting and estimation.
- **White Noise Residuals:** The error terms ϵ_t should be independent, identically distributed (i.i.d.) with mean zero and constant variance σ^2. These errors are also assumed to be uncorrelated with each other.
- **Autocorrelation Function (ACF):** For a stationary ARMA process, the *ACF will generally decay exponentially for large lags.* The rate of decay depends on the values of the AR and MA parameters.
- **Partial Autocorrelation Function (PACF):** The PACF of an ARMA model exhibits characteristic patterns. For an AR(p) process, the

DRAFT

PACF will show a sharp cutoff after lag p, while for an MA(q) process, the ACF will show a sharp cutoff after lag q.

- **Forecasting:** The ARMA model provides a framework for forecasting future values of the time series based on the past values and error terms. The forecasted values are a weighted sum of past values and past error terms.

- **Model Identification:** The correct order of the AR and MA components (i.e., p and q) can be determined by analyzing the ACF and PACF plots or by using information criteria such as AIC or BIC.

Estimating Orders p and q in ARMA(p, q) Models

Estimating the orders p and q in an ARMA(p, q) model involves determining the number of autoregressive (AR) terms and moving average (MA) terms to include. This can be done through several methods, including analyzing the ACF and PACF plots and using Information Criteria such as AIC and BIC.

1. Plotting ACF and PACF

ACF (Autocorrelation Function):

- Helps identify the MA component. For an MA(q) model, the ACF will be significant up to lag q and will drop to zero afterward.

PACF (Partial Autocorrelation Function):

- Helps identify the AR component. For an AR(p) model, the PACF will be significant up to lag p and will drop to zero afterward.

Interpretation: In the ACF plot (Figure 12.17), the significant spikes indicate potential MA terms. In the PACF plot (Figure 12.18), the significant spikes suggest potential AR terms. For example, if the ACF cuts off after lag 2 and the PACF cuts off after lag 1, an MA(2) and AR(1) model may be appropriate.

2. Information Criteria

Interpretation: In the AIC and BIC plot (Figure 12.19), the order p and q that minimize AIC and BIC are selected as the optimal values. The plot shows AIC and BIC values for different combinations of p and q, and the best model is identified by the lowest AIC or BIC value.

Python Code

The following Python code was used to estimate p and q:

DRAFT

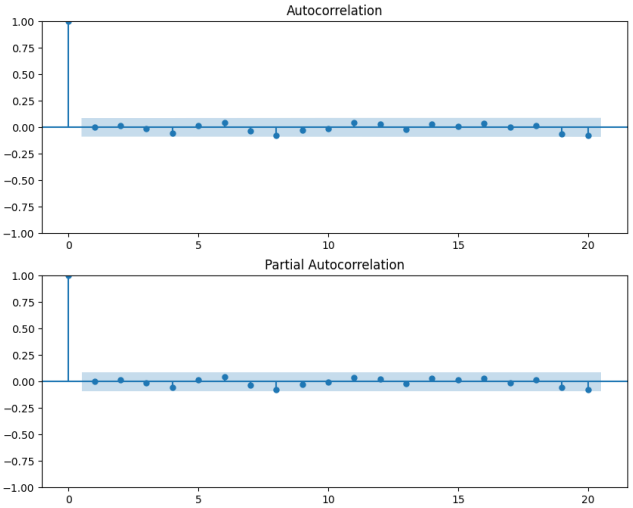


Figure 12.17: ACF and PACF plot of the residuals from a fitted ARMA model.

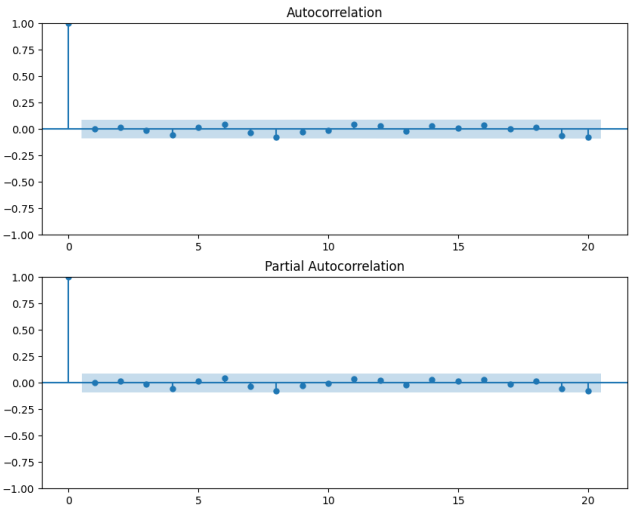


Figure 12.18: PACF plot of the residuals from a fitted ARMA model.

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_process import ArmaProcess
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf,
    plot_pacf

```

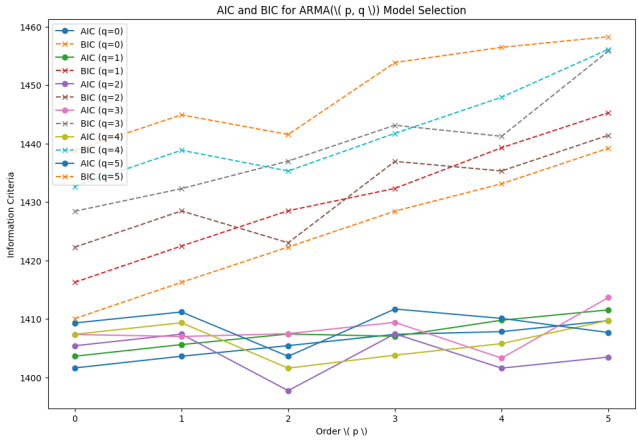


Figure 12.19: AIC and BIC values for different orders p and q.

```
# Step 1: Generate synthetic ARMA(1,1) data
# Define ARMA(1,1) model parameters
np.random.seed(42)  # For reproducibility
ar = np.array([1, -0.5])  # AR(1) with coefficient 0.5
ma = np.array([1, 0.4])   # MA(1) with coefficient 0.4

# Generate a sample time series from the ARMA process
arma_process = ArmaProcess(ar, ma)
n_samples = 500
synthetic_data = arma_process.generate_sample(nsample=
    n_samples)

# Plot the generated data
plt.figure(figsize=(10, 6))
plt.plot(synthetic_data)
plt.title('Synthetic ARMA(1,1) Time Series Data')
plt.xlabel('Time')
plt.ylabel('Value')
plt.show()

# Step 2: Plot ACF and PACF
plt.figure(figsize=(10, 6))

# Autocorrelation Function (ACF)

plot_acf(synthetic_data, lags=30, ax=plt.gca())
plt.title('Autocorrelation Function (ACF)')
plt.show()

# Partial Autocorrelation Function (PACF)
```

```
plt.figure(figsize=(10, 6))

plot_pacf(synthetic_data, lags=30, ax=plt.gca(), method='ywm')
plt.title('Partial Autocorrelation Function (PACF)')

plt.show()

# Step 3: Fit an ARMA(1,1) model to the data
# Use the ARIMA class with order=(1, 0, 1) for ARMA(1,1)
model = ARIMA(synthetic_data, order=(1, 0, 1))
fitted_model = model.fit()

# Step 4: Print the model summary
print(fitted_model.summary())

# Step 5: Forecast future values
forecast = fitted_model.forecast(steps=10)
print("Forecasted Values:", forecast)
```

Listing 12.15: Python Code for Estimating Orders p and q

12.15.7 Example: ARMA(1,1) Model

An **ARMA(1,1)** model is a specific case of the ARMA model where the autoregressive part has order 1 ($p = 1$) and the moving average part has order 1 ($q = 1$). The model is defined as:

$$y_t = \mu + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

where:

- y_t is the value of the time series at time t,
- μ is the constant term (often taken as zero for simplicity),
- ϕ_1 is the autoregressive coefficient that models the relationship between y_t and its previous value y_{t-1},
- θ_1 is the moving average coefficient that models the relationship between y_t and the previous error term ϵ_{t-1},
- ϵ_t is the white noise error term at time t.

Steps for Estimating the ARMA(1,1) Model:

- **Stationarity Check:** Ensure that the time series y_t is stationary. If it is not, you may need to difference the series to achieve stationarity.

DRAFT

- **Fit the Model:** Using statistical software or methods such as **Maximum Likelihood Estimation (MLE)**, estimate the parameters ϕ_1 and θ_1.
- **Residual Diagnostics:** After fitting the model, check the residuals ϵ_t. The residuals should be white noise, meaning they should have zero mean and constant variance, and should be uncorrelated over time.
- **Forecasting:** Once the model is fitted, you can use it to forecast future values of y_t based on the past values of the series and the estimated parameters.

12.15.8 Example: ARMA(1,1) Model

The ARMA(1,1) model is a specific case of the ARMA(p, q) model, where both the autoregressive and moving average components are of order 1. The ARMA(1,1) model is defined as:

$$y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1},$$

where:

- y_t is the value of the time series at time t,
- ϕ is the AR(1) parameter,
- θ is the MA(1) parameter,
- ϵ_t is white noise with mean 0 and variance σ^2.

Stationarity and Invertibility

- **Stationarity:** For stationarity, $|\phi| < 1$. This ensures that the effect of past values diminishes over time.
- **Invertibility:** For invertibility, $|\theta| < 1$. This ensures that the model can be expressed in terms of past values without leading to explosive behavior.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

- **ACF:** The autocorrelation function of an ARMA(1,1) model typically shows a combination of exponential decay (due to the AR component) and a spike at lag 1 (due to the MA component).
- **PACF:** The partial autocorrelation function for an ARMA(1,1) model will usually show a significant spike at lag 1, followed by a rapid tail-off.

DRAFT

Theorem 12.3 (Variance of ARMA(1,1) Model)**.** *For a stationary ARMA(1,1) process defined by $y_t = \phi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$, where $|\phi| < 1$, the variance of the process, $Var(y_t)$, is given by:*

$$Var(y_t) = \frac{1 + 2\phi\theta + \theta^2}{1 - \phi^2}\sigma^2$$

where σ^2 is the variance of the white noise process ϵ_t.

Proof. Let y_t be a stationary ARMA(1,1) process with zero mean, and let ϵ_t be a white noise process with mean zero and variance σ^2. The variance of y_t is the autocovariance at lag 0, denoted as γ_0. We derive a system of two equations for the autocovariances γ_0 and γ_1.

(i). Autocovariance at Lag 1 (γ_1): Multiply the model equation by y_{t-1} and take the expectation:

$$E[y_t y_{t-1}] = E[(\phi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1})y_{t-1}]$$

This simplifies to $\gamma_1 = \phi\gamma_0 + E[\epsilon_t y_{t-1}] + \theta E[\epsilon_{t-1}y_{t-1}]$. Given the properties of white noise and a stationary process, $E[\epsilon_t y_{t-1}] = 0$ and $E[\epsilon_{t-1}y_{t-1}] = \sigma^2$. Thus,

$$\gamma_1 = \phi\gamma_0 + \theta\sigma^2$$

(ii). Autocovariance at Lag 0 (γ_0): Multiply the model equation by y_t and take the expectation:

$$E[y_t^2] = E[(\phi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1})y_t]$$

This gives $\gamma_0 = \phi\gamma_1 + E[\epsilon_t y_t] + \theta E[\epsilon_{t-1}y_t]$. The cross-covariance terms are:

- $E[\epsilon_t y_t] = \sigma^2$
- $E[\epsilon_{t-1}y_t] = (\phi + \theta)\sigma^2$

Substituting these terms yields:

$$\gamma_0 = \phi\gamma_1 + \sigma^2(1 + \phi\theta + \theta^2)$$

(iii). Solving the System: Substitute the expression for γ_1 into the equation for γ_0:

$$\gamma_0 = \phi(\phi\gamma_0 + \theta\sigma^2) + \sigma^2(1 + \phi\theta + \theta^2)$$
$$\gamma_0(1 - \phi^2) = \sigma^2(1 + 2\phi\theta + \theta^2)$$

Solving for γ_0 gives the desired result:

$$Var(y_t) = \gamma_0 = \frac{1 + 2\phi\theta + \theta^2}{1 - \phi^2}\sigma^2$$

□

DRAFT

Theorem 12.4 (Autocorrelation Function of ARMA(1,1) Model)**.** *For a stationary ARMA(1,1) process, the autocorrelation function ρ_k is defined by:*

$$\rho_1 = \frac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2} \quad \textit{and} \quad \rho_k = \phi\rho_{k-1} \quad \textit{for } k \geq 2$$

Proof. The autocorrelation function ρ_k is defined as $\rho_k = \gamma_k/\gamma_0$.

(i). Recursive relationship for $k \geq 2$: For $k \geq 2$, the recursive relationship for the autocovariance is $\gamma_k = \phi\gamma_{k-1}$. Dividing by γ_0 gives the recursive relationship for the autocorrelation:

$$\rho_k = \phi\rho_{k-1} \quad \text{for } k \geq 2$$

(ii). Value for $k = 1$: We use the previously derived expressions for γ_0 and γ_1:

$$\gamma_0 = \frac{1+2\phi\theta+\theta^2}{1-\phi^2}\sigma^2 \quad \text{and} \quad \gamma_1 = \frac{(\phi+\theta)(1+\phi\theta)}{1-\phi^2}\sigma^2$$

The autocorrelation at lag 1 is the ratio of these two quantities:

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\frac{(\phi+\theta)(1+\phi\theta)}{1-\phi^2}\sigma^2}{\frac{1+2\phi\theta+\theta^2}{1-\phi^2}\sigma^2} = \frac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2}$$

This completes the proof. □

12.16 Models for Nonstationary Time Series and Forecasting

The ARMA(p, q) model is suitable for modeling stationary time series, where statistical properties such as mean, variance, and autocorrelation remain constant over time. However, many real-world time series exhibit nonstationary behavior, typically due to trends or varying levels. To address this, we use the Autoregressive Integrated Moving Average (ARIMA) model, which incorporates differencing to make a non-stationary series stationary before applying an ARMA model.

12.16.1 Stationarity and Differencing

A time series is said to be stationary if its statistical properties do not change over time. One common way to achieve stationarity is through differencing. The first difference of a series is given by

$$\nabla y_t = y_t - y_{t-1}$$

Higher-order differencing is denoted as:

$$\nabla^d y_t = (1-B)^d y_t$$

where B is the backshift operator such that $By_t = y_{t-1}$. Differencing helps remove trends and stabilize the mean of the series.

DRAFT

12.16.2 ARIMA(p, d, q) Model

An ARIMA(p, d, q) model is defined as an ARMA(p, q) model applied to the d-th order differenced series. The general form of the ARIMA model is given by

$$\Delta^d y_t = \phi_1 \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where,

- $\Delta^d y_t = (1 - B)^d y_t$ denotes the d-th order differenced time series,
- y_t is the observed time series at time t,
- ε_t is a white noise error term,
- $\phi_1, \ldots, \phi_p$ are the autoregressive (AR) parameters,
- $\theta_1, \ldots, \theta_q$ are the moving average (MA) parameters.

This model can also be written using the backshift operator B as

$$\phi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t,$$

where,

- $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the autoregressive polynomial,
- $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is the moving average polynomial,
- ε_t is a white noise process with mean zero and constant variance.

Special Case: ARIMA(1,1,1) Model

The ARIMA(1,1,1) model is a special case of the general ARIMA(p, d, q) model where $p = 1$, $d = 1$, and $q = 1$. It applies an ARMA(1,1) structure to the first-differenced series $\Delta y_t = y_t - y_{t-1}$. The model is written as

$$\Delta y_t = \phi_1 \Delta y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

where,

- $\Delta y_t = y_t - y_{t-1}$ is the first difference of the time series,
- ϕ_1 is the autoregressive coefficient,
- θ_1 is the moving average coefficient,
- ε_t is white noise.

DRAFT

Alternatively, the ARIMA(1,1,1) model can be expressed using the backshift operator B as

$$(1 - \phi_1 B)(1 - B)y_t = (1 + \theta_1 B)\varepsilon_t.$$

Expanding this equation, we have

$$y_t - y_{t-1} - \phi_1(y_{t-1} - y_{t-2}) = \varepsilon_t + \theta_1 \varepsilon_{t-1},$$

which highlights how both the lagged differences and past error terms contribute to the value of y_t. This simple ARIMA(1,1,1) model is widely used in practice and often provides a good fit for time series with a linear trend and short memory.

Remark 12.16.1. *In some ARIMA models, a constant term μ may be included to capture a possible drift in the differenced series. However, if the differenced series has zero mean, μ can be omitted. In this chapter, we consider models without a drift term unless stated otherwise.*

12.16.3 Model Identification

Model identification is the critical first step in fitting an ARIMA(p, d, q) model, involving the selection of appropriate orders p, d, and q based on the observed time series characteristics.

Determining the Differencing Order d

The differencing order d is chosen to achieve stationarity in the time series. The main approaches are

- **Visual Inspection:** Plot the original time series and look for trends or changing variance that suggest non-stationarity.
- **Unit Root Tests:** Statistical tests that formally assess stationarity.
  - **Augmented Dickey-Fuller (ADF) Test:** Tests the null hypothesis of a unit root (non-stationarity). A significant test statistic indicates stationarity.
  - **KPSS Test:** Tests the null hypothesis of stationarity. A significant result suggests non-stationarity.
- **Differencing:** Apply differencing one or more times and re-examine the series until stationarity is achieved.

DRAFT

Choosing the AR order p and MA order q

Once stationarity is attained, the orders p and q are identified by examining the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Common patterns for model identification include:

- **AR(p) process:** PACF cuts off after lag p, ACF tails off.
- **MA(q) process:** ACF cuts off after lag q, PACF tails off.
- **ARMA(p, q) process:** Both ACF and PACF tail off gradually.

Information Criteria

To objectively compare candidate models, information criteria such as

- **Akaike Information Criterion (AIC):**

$$\text{AIC} = -2 \log \widehat{L} + 2k,$$

  where $\widehat{L}$ is the maximum likelihood of the model, evaluated at the estimated and k is the number of estimated parameters.

- **Bayesian Information Criterion (BIC):**

$$\text{BIC} = -2 \log \widehat{L} + k \log n,$$

  where n is the sample size.

Models with lower AIC or BIC values are preferred.

12.16.4 Estimation of Parameters in ARIMA Models

Once an appropriate ARIMA(p, d, q) model is identified, the next step is to estimate its parameters. The parameters of the model are

- Autoregressive parameters: $\phi_1, \phi_2, \ldots, \phi_p$
- Moving average parameters: $\theta_1, \theta_2, \ldots, \theta_q$
- (Optional) Drift term or constant: μ
- The variance of the white noise process: σ^2

There are several methods available for parameter estimation in ARIMA models, the most commonly used being:

DRAFT

1. Maximum Likelihood Estimation (MLE) This is the most widely used and statistically robust method for estimating ARIMA parameters. MLE finds the parameter values that maximize the likelihood of observing the given time series data under the specified model. It assumes that the innovations, or error terms ε_t, are independently and identically distributed following a normal distribution, i.e., $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

The process involves constructing a likelihood function from the model's structure and then using numerical optimization algorithms to find the parameters that maximize this function. The resulting estimates are asymptotically efficient and normally distributed, which allows for the calculation of confidence intervals and hypothesis testing.

For the ARIMA(p, d, q) model, the log-likelihood function for the differenced series can be expressed as:

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

where T is the length of the differenced series, and the innovations ε_t are calculated recursively.

2. Conditional Sum of Squares (CSS) Conditional Sum of Squares is an alternative method that is computationally simpler than MLE. It is based on minimizing the sum of squared errors of the model. The objective function to be minimized is:

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=q+1}^{T}\varepsilon_t^2$$

This method requires making an assumption about the initial values of the innovations $(\varepsilon_1, \ldots, \varepsilon_q)$, such as setting them to zero. While CSS is faster, its estimates are generally less efficient than those from MLE, especially for models with a strong moving average component. It is often used to provide initial parameter guesses for the more complex MLE optimization process.

3. Yule-Walker Equations This method is a non-iterative approach that is only applicable to pure autoregressive (AR) models (i.e., when $q = 0$). The Yule-Walker equations relate the AR coefficients $(\phi_1, \phi_2, \ldots, \phi_p)$ to the sample autocorrelations (r_k) of the time series. This method involves solving a system of linear equations to find the parameters:

DRAFT

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}$$

The Yule-Walker method provides direct estimates for the AR parameters but is not suitable for models that include a moving average component. The estimates are also less efficient than those obtained through MLE.

12.16.5 Forecasting Equation of ARIMA(p, d, q) Model

Once an ARIMA(p, d, q) model is fitted, forecasting involves predicting future values of the time series using the estimated parameters and past observations. For 1-step ahead forecasting, the equation is

$$\widehat{y}_{t+1} = y_t + \widehat{\Delta^d y}_{t+1},$$

where,

$$\widehat{\Delta^d y}_{t+1} = \widehat{\phi}_1 \Delta^d y_t + \cdots + \widehat{\phi}_p \Delta^d y_{t+1-p} + \widehat{\theta}_1 \varepsilon_t + \cdots + \widehat{\theta}_q \varepsilon_{t+1-q}.$$

For h-step ahead forecasts ($h \geq 2$), the forecast is generated recursively by replacing future values with their own forecasts, and assuming that future error terms are zero:

$$\widehat{\varepsilon}_{t+h} = 0 \quad \text{for } h \geq 1.$$

The final forecast of y_{t+h} is obtained by inverting the differencing:

$$\widehat{y}_{t+h} = \widehat{y}_{t+h-1} + \widehat{\Delta y}_{t+h}.$$

Forecasting Equation for the ARIMA(1,1,1) Model

After fitting an ARIMA(1,1,1) model,

$$\Delta y_t = \phi_1 \Delta y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

the 1-step ahead forecast for the differenced series Δy_{t+1} is given by

$$\widehat{\Delta y}_{t+1} = \widehat{\phi}_1 \Delta y_t + \widehat{\theta}_1 \varepsilon_t,$$

where ε_t are the residuals estimated from the fitted model.

The forecast for the original series y_{t+1} is then obtained by inverting the differencing:

$$\widehat{y}_{t+1} = y_t + \widehat{\Delta y}_{t+1}.$$

DRAFT

For multi-step forecasts ($h \geq 2$), since future error terms are unknown, we assume

$$\widehat{\varepsilon}_{t+h} = 0,$$

and recursively compute

$$\widehat{\Delta y}_{t+h} = \widehat{\phi}_1 \widehat{\Delta y}_{t+h-1},$$

leading to

$$\widehat{y}_{t+h} = \widehat{y}_{t+h-1} + \widehat{\Delta y}_{t+h}.$$

12.16.6 Model Diagnostics

After fitting an ARIMA model and estimating its parameters, the next crucial step is to perform model diagnostics. The primary objective of this process is to assess whether the model is a good fit for the data and whether the underlying assumptions have been satisfied. A well-specified ARIMA model should have residuals that resemble a white noise process—that is, the residuals should have a mean of zero, constant variance, and no autocorrelation.

1. Analysis of Residuals

Residuals are the differences between the observed data and the values predicted by the model. If the model is correctly specified, the residuals should not contain any systematic information. In other words, they should be uncorrelated and normally distributed. The following diagnostic tools are commonly used:

- **Residual Plot:** A time series plot of the residuals should display a random scatter around zero. Any visible patterns such as trends, cycles, or changes in variance indicate that the model has not fully captured the underlying structure of the data.
- **ACF and PACF of Residuals:** The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the residuals should not exhibit any significant autocorrelation. All autocorrelation coefficients should lie within the confidence bounds. Significant spikes in these plots suggest model misspecification.
- **Normality of Residuals:** Residuals should follow a normal distribution. This can be assessed using:
  - A histogram of residuals — should approximate a bell-shaped curve.
  - A Q-Q plot — the residuals should lie close to the 45-degree line.

DRAFT

2. Formal Statistical Tests

In addition to graphical diagnostics, formal statistical tests provide rigorous assessment of the residuals:

- **Ljung-Box Test:** This test evaluates whether the residuals are independently distributed (i.e., resemble white noise). The null hypothesis H_0 is that the residuals are uncorrelated. A high p-value (greater than a chosen significance level such as 0.05) indicates that the residuals are consistent with white noise. A low p-value suggests that the model may not have adequately captured all serial dependence in the data.
- **Augmented Dickey-Fuller (ADF) Test:** Although typically used to test the stationarity of the original series, the ADF test can also be applied to residuals to ensure they are stationary.

If the diagnostic checks reveal problems—such as significant autocorrelations or non-normal residuals—the ARIMA model may be inadequate. In such cases, it is necessary to return to the model identification step, modify the parameters p, d, or q, and re-estimate the model until residuals behave like white noise.

Fitting ARIMA Model in Python

We can fit an ARIMA(p, d, q) model in Python using the `statsmodels` library. Below is a step-by-step example:

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf,
  plot_pacf

# Load a time series dataset (for example, airline
  passengers)
from statsmodels.datasets import get_rdataset
data = get_rdataset("AirPassengers").data
ts = data["value"]
ts.index = pd.date_range(start="1949-01", periods=len(ts),
   freq="M")

# Visualize the data
ts.plot(title="Monthly Airline Passenger Numbers")
plt.show()

# Fit ARIMA(p,d,q) model (e.g., ARIMA(1,1,1))
model = ARIMA(ts, order=(1, 1, 1))
fitted_model = model.fit()

```

DRAFT

```
# Print the model summary
print(fitted_model.summary())

# Forecast the next 12 months
forecast = fitted_model.get_forecast(steps=12)
forecast_mean = forecast.predicted_mean
conf_int = forecast.conf_int()

# Plot the forecast with confidence intervals
plt.figure(figsize=(10, 5))
plt.plot(ts, label="Observed")
plt.plot(forecast_mean.index, forecast_mean, label="
  Forecast", color="red")
plt.fill_between(forecast_mean.index, conf_int.iloc[:, 0],
   conf_int.iloc[:, 1], color="pink", alpha=0.3)
plt.title("ARIMA Forecast")
plt.legend()
plt.show()
```

Listing 12.16: Fitting ARIMA Model in Python

12.17 The Seasonal ARIMA Model

While the ARIMA model is a powerful tool for analyzing time series data, it is not always effective for data that exhibits a clear and repeating seasonal pattern. Seasonal patterns are cyclical fluctuations that occur at a fixed and known frequency, such as monthly sales data that peaks every December, or quarterly revenue data with a consistent Q4 surge. To capture these regular, periodic movements, we extend the ARIMA framework with seasonal components, creating the Seasonal Autoregressive Integrated Moving Average (SARIMA) model.

The SARIMA model combines the non-seasonal ARIMA components with their seasonal counterparts. The full model is denoted by the notation

$$\text{SARIMA}(p, d, q)(P, D, Q)_s.$$

- The first set of parameters, (p, d, q), represents the **non-seasonal** part of the model, which you are already familiar with.
- The second set of parameters, $(P, D, Q)_s$, represents the **seasonal** part of the model.
  - P: The number of seasonal autoregressive (SAR) terms.
  - D: The number of seasonal differences.
  - Q: The number of seasonal moving average (SMA) terms.
  - s: The length of the seasonal period (e.g., $s = 12$ for monthly data with a yearly cycle, or $s = 4$ for quarterly data).

DRAFT

12.17.1 The SARIMA Model Equation

To express the SARIMA model mathematically, we introduce a new backshift operator, B^s, which shifts the time series by s periods.

$$B^s y_t = y_{t-s}$$

Using this operator, we can define the seasonal differencing and seasonal AR and MA terms. The complete SARIMA model equation can be written as

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP})\Delta^d \Delta_s^D y_t \\ = (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^s + \cdots + \Theta_Q B^{sQ})\varepsilon_t$$

where,

- $\Delta^d y_t = (1-B)^d y_t$ is the non-seasonal d-th order differenced time series, consistent with the ARIMA model.
- $\Delta_s^D y_t = (1-B^s)^D y_t$ is the seasonal D-th order differenced time series.
- y_t is the observed time series at time t.
- ε_t is a white noise error term.
- $\phi_1, \ldots, \phi_p$ are the non-seasonal AR parameters.
- $\theta_1, \ldots, \theta_q$ are the non-seasonal MA parameters.
- $\Phi_1, \ldots, \Phi_P$ are the seasonal AR parameters.
- $\Theta_1, \ldots, \Theta_Q$ are the seasonal MA parameters.

This equation shows that the SARIMA model factors the non-seasonal and seasonal components together. For instance, the seasonal AR terms at time t are multiplied by the non-seasonal AR terms. This allows the model to simultaneously account for both short-term, non-seasonal dependencies and long-term, seasonal dependencies.

12.17.2 Example: A SARIMA(0, 1, 1)(0, 1, 1)$_{12}$ Model

For SARIMA$(0,1,1)(0,1,1)_{12}$ we have no AR terms (nonseasonal or seasonal), one nonseasonal MA parameter θ, one seasonal MA parameter Θ, first nonseasonal difference $d = 1$, and first seasonal difference $D = 1$ with $s = 12$. Using the backshift operator B, we have

$$(1)(1)\,\Delta^1 \Delta_{12}^1 y_t = (1 + \theta B)\,(1 + \Theta B^{12})\,\varepsilon_t,$$

so

$$\Delta\Delta_{12} y_t = (1 + \theta B + \Theta B^{12} + \theta\Theta B^{13})\,\varepsilon_t.$$

DRAFT

The left side of the equation represents the double differencing (non-seasonal and seasonal) of the time series. The right side shows how this differenced series is modeled by both non-seasonal and seasonal moving average components. This model structure is often effective for forecasting monthly retail sales or electricity consumption, where both an overall upward trend and a repeating annual cycle are present.

Model (expanded in y_t and ε_t)

To better understand the components, we can expand the backshift terms. The differencing on the left-hand side can be written as:

$$\begin{aligned}(1-B)(1-B^{12})y_t &= (1-B-B^{12}+B^{13})y_t \\ &= y_t - y_{t-1} - y_{t-12} + y_{t-13} \\ &= (y_t - y_{t-1}) - (y_{t-12} - y_{t-13})\end{aligned}$$

Therefore the model written in observation/error form is

$$y_t - y_{t-1} - y_{t-12} + y_{t-13} = \varepsilon_t + \theta_1\varepsilon_{t-1} + \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13}$$

This form shows the differenced observations on the left and their moving average representation of current and past innovations on the right.

12.17.3 Example: A SARIMA$(1,1,1)(1,1,1)_{12}$ Model

For SARIMA$(1,1,1)(1,1,1)_{12}$ we have one nonseasonal AR parameter ϕ, one nonseasonal MA parameter θ, one seasonal AR parameter Φ, one seasonal MA parameter Θ, first nonseasonal difference $d = 1$, and first seasonal difference $D = 1$ with $s = 12$. Using the backshift operator B:

$$(1-\phi B)\,(1-\Phi B^{12})\,\Delta^1\Delta_{12}^1 y_t = (1+\theta B)\,(1+\Theta B^{12})\,\varepsilon_t,$$

so

$$(1-\phi B-\Phi B^{12}+\phi\Phi B^{13})\,\Delta\Delta_{12} y_t = (1+\theta B+\Theta B^{12}+\theta\Theta B^{13})\,\varepsilon_t.$$

Model (expanded in y_t and ε_t)

First evaluate the left differencing operator, we have

$$(1-B)(1-B^{12})y_t = y_t - y_{t-1} - y_{t-12} + y_{t-13}.$$

Substituting this into the AR and MA parts, the full model can then be written as

$$\begin{aligned}(1-\phi_1 B-\Phi_1 B^{12}+\phi_1\Phi_1 B^{13})&(y_t - y_{t-1} - y_{t-12} + y_{t-13}) \\ &= (1+\theta_1 B+\Theta_1 B^{12}+\theta_1\Theta_1 B^{13})\varepsilon_t.\end{aligned}$$

DRAFT

Thus,

$$\begin{aligned}\big[y_t - y_{t-1} - y_{t-12} + y_{t-13} - \phi(y_{t-1} - y_{t-2} - y_{t-13} + y_{t-14}) \\ - \Phi(y_{t-12} - y_{t-13} - y_{t-24} + y_{t-25}) + \phi\Phi(y_{t-13} - y_{t-14} - y_{t-25} + y_{t-26})\big] \\ = \varepsilon_t + \theta\,\varepsilon_{t-1} + \Theta\,\varepsilon_{t-12} + \theta\Theta\,\varepsilon_{t-13}.\end{aligned}$$

12.17.4 Parameter Estimation

Estimating the parameters for a SARIMA model is typically done in three main steps: model identification, parameter estimation, and diagnostic checking. This process is iterative, and you may need to go back and forth between the steps to find the best model.

1. Model Identification

Selecting appropriate orders for the SARIMA model, $(p, d, q)(P, D, Q)_s$ components is a crucial step to ensure a good fit and accurate forecasting. This involves determining the nonseasonal orders (p, d, q) and seasonal orders (P, D, Q) based on the observed data patterns.

(i). Selection of Differencing Orders d and D

- **Seasonal differencing order D:** Begin by examining the time series for seasonal patterns with period s. If a strong seasonal pattern is observed (e.g., spikes at lag s in the autocorrelation function), apply seasonal differencing first:

$$\Delta_s^D y_t = (1 - B^s)^D y_t,$$

  where typically $D = 0$ or 1.

- **Nonseasonal differencing order d:** After removing seasonality, check for remaining trends or unit roots in the seasonally differenced series. Apply nonseasonal differencing as needed:

$$\Delta^d y_t = (1 - B)^d y_t,$$

  where d is usually 0 or 1.

- **Order of differencing:** Seasonal differencing should be applied *before* nonseasonal differencing to properly isolate seasonal effects.

(ii). Selection of AR Orders p and P

- **Nonseasonal AR order p:** After differencing, analyze the partial autocorrelation function (PACF) of the transformed series at low lags (e.g., lag 1, 2, 3, ...). A significant spike at lag p followed by a cutoff suggests an AR(p) process.

DRAFT

- **Seasonal AR order P:** Inspect the PACF at seasonal lags (multiples of s, e.g., s, $2s$, $3s$, ...). A significant spike at lag Ps indicates the presence of a seasonal AR(P) component.

(iii). Selection of MA Orders q and Q

- **Nonseasonal MA order q:** Examine the autocorrelation function (ACF) of the differenced series at low lags. A sharp cutoff after lag q suggests an MA(q) component.
- **Seasonal MA order Q:** Look at the ACF at seasonal lags Qs. A significant spike at lag Qs implies a seasonal MA(Q) component.

Practical Tips

- Use plots of ACF and PACF on appropriately differenced data to guide initial order selection.
- Typically, start with small orders ($p, q, P, Q \leq 2$) and increase if diagnostics suggest inadequacy.
- Information criteria such as AIC or BIC can help compare competing models.
- Residual diagnostics are essential to validate model adequacy.

2. Parameter Estimation

Once a candidate model order is chosen, the parameters

$$\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q$$

are estimated. The most widely used method for this is Maximum Likelihood Estimation (MLE).

Maximum Likelihood Estimation (MLE)

MLE is an optimization technique that finds the parameter values that maximize the likelihood of the observed data occurring under the model. Assuming the innovations, ε_t, are normally distributed white noise, the goal is to maximize the log-likelihood function, ℓ, to obtain the parameter estimates, $\widehat{\boldsymbol{\theta}}$.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|y_1, \ldots, y_T)$$

3. Diagnostic Checking

The final step is to validate the fitted model by checking if its assumptions are met. This is primarily done by analyzing the model's residuals.

DRAFT

Residual Analysis

A good SARIMA model should produce residuals that are uncorrelated, have a zero mean, and a constant variance. In other words, the residuals should be **white noise**.

- **Residual Plots:** Visual inspection of the residuals over time helps to identify any remaining trends, seasonality, or changes in variance.
- **ACF of Residuals:** An ACF plot of the residuals should show no significant spikes, indicating that no systematic information has been left in the residuals.
- **Ljung-Box Test:** This formal statistical test checks for the presence of significant autocorrelation in the residuals. A non-significant p-value (e.g., $p > 0.05$) is desired, as it suggests the residuals are indeed uncorrelated. If the test rejects the null hypothesis of no autocorrelation, the model may be inadequate and the process must be repeated.

12.17.5 Python Code: SARIMA$(p, d, q)(P, D, Q)_s$ Model

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller

# Load your time series data as a pandas Series 'y'
# y = pd.read_csv('your_timeseries.csv', index_col=0, parse_dates=True)['value']

# Plot original series
plt.figure(figsize=(10,4))
plt.plot(y)
plt.title('Original Time Series')
plt.show()

# Seasonal decomposition to visualize seasonal pattern
decomp = seasonal_decompose(y, model='additive', period=12)
decomp.plot()
plt.show()

# Check stationarity with Augmented Dickey-Fuller test
adf_result = adfuller(y.dropna())
print(f'ADF Statistic: {adf_result[0]:.4f}, p-value: {adf_result[1]:.4f}')
```

```
25
26  # Apply seasonal differencing if strong seasonality
      present (e.g., s=12)
27  y_seasonal_diff = y.diff(12).dropna()
28  plt.plot(y_seasonal_diff)
29  plt.title('Seasonally Differenced Series (lag 12)')
30  plt.show()
31
32  # Test stationarity again after seasonal differencing
33  adf_seasonal = adfuller(y_seasonal_diff)
34  print(f'Seasonal Diff ADF Statistic: {adf_seasonal[0]:.4f
      }, p-value: {adf_seasonal[1]:.4f}')
35
36  # Apply nonseasonal differencing if needed
37  y_diff = y_seasonal_diff.diff().dropna()
38  plt.plot(y_diff)
39  plt.title('Seasonally and Nonseasonally Differenced Series
      ')
40  plt.show()
41
42  # Check stationarity again
43  adf_final = adfuller(y_diff)
44  print(f'Final Differenced ADF Statistic: {adf_final[0]:.4f
      }, p-value: {adf_final[1]:.4f}')
45
46  # Plot ACF and PACF of the final differenced series to
      identify p, q, P, Q
47  fig, axes = plt.subplots(2, 1, figsize=(12,8))
48  plot_acf(y_diff, lags=40, ax=axes[0])
49  axes[0].set_title('ACF Plot')
50  plot_pacf(y_diff, lags=40, ax=axes[1])
51  axes[1].set_title('PACF Plot')
52  plt.tight_layout()
53  plt.show()
54
55  # Interpretation guide:
56  print("""
57  - Look for significant spikes at low lags in PACF to
      suggest nonseasonal AR order p.
58  - Look for significant spikes at seasonal lags (e.g., 12,
      24) in PACF to suggest seasonal AR order P.
59  - Look for significant spikes at low lags in ACF to
      suggest nonseasonal MA order q.
60  - Look for significant spikes at seasonal lags in ACF to
      suggest seasonal MA order Q.
61  - Differencing orders d and D chosen based on stationarity
       tests and plots above.
62  """)
```

Listing 12.17: Identifying SARIMA orders using ACF, PACF, and differencing

12.18 Exponential Smoothing and the ETS Framework

Smoothing techniques are essential for analyzing time series data, as they help to reduce noise and highlight underlying trends and patterns. Moving Average also a smoothig technique that explained in Section 12.7.3. Exponential smoothing is a family of time-series forecasting methods that *assign exponentially decreasing weights to past observations.* The core idea is that recent observations are more relevant for predicting the future than older ones. These methods are intuitive, flexible, and have been widely used in a variety of fields, including business, economics, and engineering.

Over time, different exponential smoothing methods were developed to handle various time-series components, such as trend and seasonality. This led to a need for a unified framework to classify these models. The ETS (Error, Trend, Seasonality) framework provides a systematic way to categorize and select the appropriate exponential smoothing model for a given time series. This framework classifies models based on the nature of their error, trend, and seasonality components, each of which can be either additive (A) or multiplicative (M), or none (N).

This section will introduce the fundamental concepts of exponential smoothing, starting with the simplest model, Simple Exponential Smoothing (SES). We will then explore more advanced models, such as Holt's Linear Trend Model and the Holt-Winters Seasonal Model, and demonstrate how they fit within the comprehensive ETS framework. We will also provide the underlying mathematical equations and practical Python code examples to help you understand and implement these powerful forecasting techniques.

12.18.1 Exponential Smoothing

The simplest form of the technique, known as **Simple Exponential Smoothing**, is used for time series data that does not have a significant trend or seasonal pattern. The forecasting value is a weighted average of the most recent actual observation and the previous. It is calculated by using following equation.

- Simple Exponential Smoothing (SES)
- Holt's Linear Trend Model
- Holt-Winters Seasonal Model
  - Additive Seasonality
  - Multiplicative Seasonality

DRAFT

12.18.2 Simple Exponential Smoothing

Simple Exponential Smoothing (SES) is used for time series data that does not exhibit a trend or seasonality. It smooths the data using an exponentially weighted moving average.

$$\widehat{y}_{t+1} = \alpha y_t + (1 - \alpha)\widehat{y}_t$$

where:

- $\widehat{y}_{t+1}$ is the forecasted value for the next time period,
- y_t is the actual value at time t,
- $\widehat{y}_t$ is the smoothed value at time t,
- α is the smoothing factor $(0 < \alpha < 1)$.

Python Code for SES

```
from statsmodels.tsa.holtwinters import SimpleExpSmoothing

# Fit Simple Exponential Smoothing model
model_ses = SimpleExpSmoothing(data['Passengers']).fit(
  smoothing_level=0.2, optimized=False)

# Forecast future values
data['SES'] = model_ses.fittedvalues

# Plot the original data and SES
plt.figure(figsize=(12, 6))
plt.plot(data['Passengers'], label='Original Data')
plt.plot(data['SES'], label='SES', color='purple')
plt.title('Simple Exponential Smoothing')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()
```

Listing 12.18: Calcualting and Ploting Simple Exponential Smoothing Curve

The simple exponential smoothing line is presented in Figure 12.20.

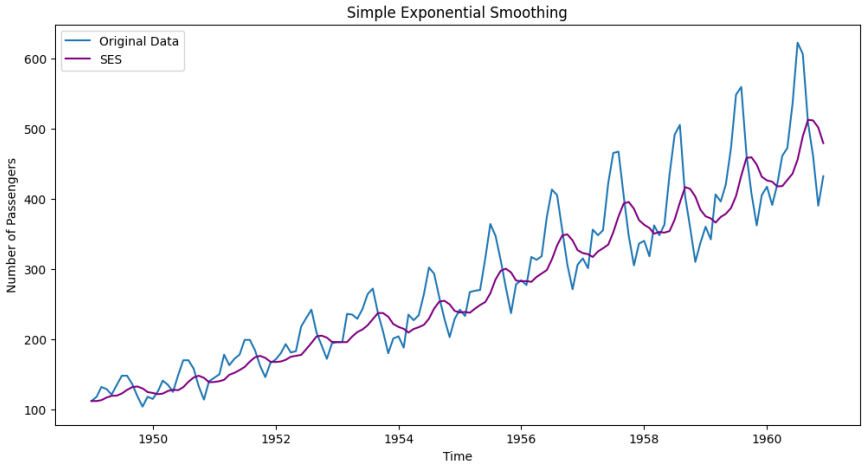


Figure 12.20: Weighted Moving Average of 12 Months.

The Smoothing Constant (α)

The smoothing constant, α, is the most important parameter in exponential smoothing. It dictates how much weight is given to the most recent observation:

- When α is **high** (close to 1), the forecast places a strong emphasis on the most recent data.
- When α is **low** (close to 0), the forecast gives more weight to past forecasts, resulting in a smoother trend line.

To begin the forecasting process, you need an initial forecast ($\widehat{y}_1$), commonly set as the first actual observation (y_1).

Problem 12.6. *A company is forecasting its monthly sales using Simple Exponential Smoothing. The actual sales for the first five months are given below.*

Month	***Actual Sales*** (y_t)
1	*100*
2	*110*
3	*125*
4	*115*
5	*130*

Using an initial forecast of $\widehat{y}_1 = 100$ *and a smoothing constant of* $\alpha = 0.2$*, calculate the exponential smoothing forecast for each month and the final forecast for Month 6.*

DRAFT

Solution: The formula for Simple Exponential Smoothing is

$$\widehat{y}_{t+1} = \alpha y_t + (1 - \alpha)\widehat{y}_t$$

Given $\alpha = 0.2$, this becomes:

$$\widehat{y}_{t+1} = 0.2y_t + 0.8\widehat{y}_t$$

We calculate the forecast for each month iteratively.

- **Forecast for Month 2 ($\widehat{y}_2$):** Using the initial forecast $\widehat{y}_1 = 100$ and actual sales $y_1 = 100$. $\widehat{y}_2 = (0.2)(100) + (0.8)(100) = 20 + 80 = 100$
- **Forecast for Month 3 ($\widehat{y}_3$):** Using $\widehat{y}_2 = 100$ and actual sales $y_2 = 110$. $\widehat{y}_3 = (0.2)(110) + (0.8)(100) = 22 + 80 = 102$
- **Forecast for Month 4 ($\widehat{y}_4$):** Using $\widehat{y}_3 = 102$ and actual sales $y_3 = 125$. $\widehat{y}_4 = (0.2)(125) + (0.8)(102) = 25 + 81.6 = 106.6$
- **Forecast for Month 5 ($\widehat{y}_5$):** Using $\widehat{y}_4 = 106.6$ and actual sales $y_4 = 115$. $\widehat{y}_5 = (0.2)(115) + (0.8)(106.6) = 23 + 85.28 = 108.28$
- **Forecast for Month 6 ($\widehat{y}_6$):** Using $\widehat{y}_5 = 108.28$ and actual sales $y_5 = 130$. $\widehat{y}_6 = (0.2)(130) + (0.8)(108.28) = 26 + 86.624 = 112.624$

The complete results are summarized in the table below.

Month	**Actual Sales (y_t)**	**Forecast ($\widehat{y}_t$)**
1	100	100 (Initial)
2	110	100
3	125	102
4	115	106.6
5	130	108.28
6	-	**112.624**

The forecast for Month 6 is 112.624.

Advantages of SES

- **Extremely simple:** Easy to understand and implement, making it a good baseline model.
- **Minimal data requirements:** Only one smoothing constant (α) needs to be chosen.
- **Works well for stable data:** It is highly effective for time series that fluctuate around a constant mean.

DRAFT

Disadvantages

- **Cannot handle trends:** It will always produce a flat forecast that lags behind a series with an upward or downward trend.
- **Cannot handle seasonality:** Like trends, any seasonal pattern will not be captured in the forecast. It is only suitable for non-seasonal, non-trending data.

12.18.3 Holt's Linear Trend Model

Holt's Linear Trend Model extends SES by incorporating trends. It smooths the data using two components: the level and the trend.

$$\begin{aligned} \text{Level:} \quad & L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \\ \text{Trend:} \quad & T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ \text{Forecast:} \quad & \widehat{y}_{t+h} = L_t + hT_t \end{aligned}$$

where,

- L_t is the level at time t,
- T_t is the trend at time t,
- α is the smoothing parameter for the level,
- β is the smoothing parameter for the trend,
- h is the forecast horizon.

Problem 12.7. *A company is forecasting its monthly sales using Holt's Linear Trend Model. The actual sales for the first five months are provided in the table below.*

Month	***Actual Sales*** (y_t)
1	*100*
2	*110*
3	*125*
4	*115*
5	*130*

Using initial values of $L_1 = 100$ *and* $T_1 = 10$*, a level smoothing constant of* $\alpha = 0.2$*, and a trend smoothing constant of* $\beta = 0.3$*, calculate the Holt's linear trend forecast for each month and the final forecast for Month 6.*

DRAFT

Solution: Given $\alpha = 0.2$ and $\beta = 0.3$, Holt's Linear Trend Model equations become

$$L_t = 0.2y_t + 0.8(L_{t-1} + T_{t-1})$$
$$T_t = 0.3(L_t - L_{t-1}) + 0.7T_{t-1}$$

We are given initial values $L_1 = 100$ and $T_1 = 10$. We will calculate the values iteratively.

- **For Month 2:** $L_2 = (0.2)(110) + (0.8)(100 + 10) = 22 + 88 = 110$ $T_2 = (0.3)(110 - 100) + (0.7)(10) = (0.3)(10) + 7 = 3 + 7 = 10$ $\widehat{y}_3 = L_2 + 1T_2 = 110 + 10 = 120$
- **For Month 3:** $L_3 = (0.2)(125) + (0.8)(110 + 10) = 25 + 96 = 121$ $T_3 = (0.3)(121 - 110) + (0.7)(10) = (0.3)(11) + 7 = 3.3 + 7 = 10.3$ $\widehat{y}_4 = L_3 + 1T_3 = 121 + 10.3 = 131.3$
- **For Month 4:** $L_4 = (0.2)(115)+(0.8)(121+10.3) = 23+(0.8)(131.3) = 23+105.04 = 128.04$ $T_4 = (0.3)(128.04-121)+(0.7)(10.3) = (0.3)(7.04)+7.21 = 2.112 + 7.21 = 9.322$ $\widehat{y}_5 = L_4 + 1T_4 = 128.04 + 9.322 = 137.362$
- **For Month 5:** $L_5 = (0.2)(130) + (0.8)(128.04 + 9.322) = 26 + (0.8)(137.362) = 26 + 109.8896 = 135.8896$ $T_5 = (0.3)(135.8896 - 128.04) + (0.7)(9.322) = (0.3)(7.8496) + 6.5254 = 2.35488 + 6.5254 = 8.88028$ $\widehat{y}_6 = L_5 + 1T_5 = 135.8896 + 8.88028 = 144.76988$

The forecast for Month 6 is approximately **144.77**. The complete results are summarized in the table below.

Month	**Actual Sales (y_t)**	**Level (L_t)**	**Trend (T_t)**	**Forecast ($\widehat{y}_{t+1}$)**
1	100	100	10	110
2	110	110	10	120
3	125	121	10.3	131.3
4	115	128.04	9.322	137.362
5	130	135.8896	8.88028	144.76988
6	-	-	-	**144.77**

Advantages of Holt's Linear Trend Model

- **Captures linear trends:** It is effective for time series that show a consistent, linear increase or decrease over time.
- **More accurate than SES for trending data:** By incorporating a trend component, it provides a much more accurate forecast for trending series.

DRAFT

Disadvantages of Holt's Linear Trend Model

- **Cannot handle seasonality:** Like SES, it fails to capture repeating seasonal patterns.
- **Over-forecasting:** The linear trend component can sometimes produce unrealistic forecasts that trend upward or downward indefinitely, especially for long-term predictions. This issue is often addressed with dampened trend models.

Python Code for Holt's Linear Trend Model

```
from statsmodels.tsa.holtwinters import Holt

# Fit Holts Linear Trend model
model_holt = Holt(data['Passengers']).fit(smoothing_level
  =0.8, smoothing_slope=0.2, optimized=False)

# Forecast future values
data['Holt'] = model_holt.fittedvalues

# Plot the original data and Holts model
plt.figure(figsize=(12, 6))
plt.plot(data['Passengers'], label='Original Data')
plt.plot(data['Holt'], label='Holts Model', color='orange'
  )
plt.title('Holts Linear Trend Model')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()
```

12.18.4 Holt-Winters Seasonal Model

The Holt-Winters Seasonal Model extends Holt's model to account for seasonality. It includes seasonal components in addition to the level and trend.

Additive Seasonality

For additive seasonality:

$$\begin{aligned} \text{Level:} \quad & L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1}) \\ \text{Trend:} \quad & T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ \text{Seasonality:} \quad & S_t = \gamma(y_t - L_{t-1} - T_{t-1}) + (1-\gamma)S_{t-s} \\ \text{Forecast:} \quad & \widehat{y}_{t+h} = L_t + hT_t + S_{t+h-s} \end{aligned}$$

where,

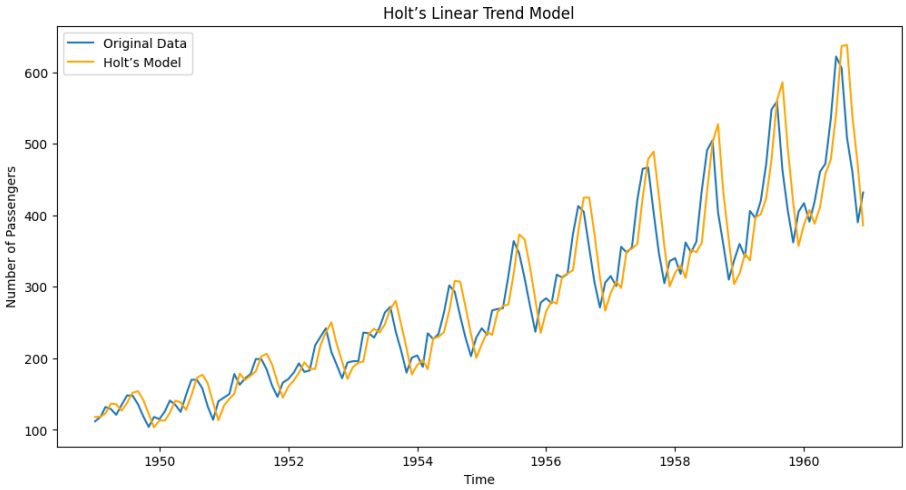


Figure 12.21: Holt's Linear Trend Model of Air Passengers Data

- S_t is the seasonal component,
- γ is the smoothing parameter for seasonality,
- s is the length of the seasonal cycle.

Multiplicative Seasonality For multiplicative seasonality:

$$\begin{aligned}
\text{Level:} \quad & L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1}) \\
\text{Trend:} \quad & T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\
\text{Seasonality:} \quad & S_t = \gamma \frac{y_t}{L_{t-1} + T_{t-1}} + (1-\gamma)S_{t-s} \\
\text{Forecast:} \quad & \widehat{y}_{t+h} = (L_t + hT_t)S_{t+h-s}
\end{aligned}$$

where,

- S_t is the seasonal component (multiplied),
- γ is the smoothing parameter for seasonality,
- s is the length of the seasonal cycle.

Problem 12.8. *A company is forecasting its quarterly sales using a Holt-Winters Additive Seasonal Model. The sales data for the first two years (8 quarters) are given below.*

DRAFT

Quarter	***Year 1 (y_t)***	***Year 2 (y_t)***
Q1	*100*	*110*
Q2	*120*	*135*
Q3	*150*	*160*
Q4	*110*	*125*

Using the following initial values, and smoothing constants $\alpha = 0.2$, $\beta = 0.3$, and $\gamma = 0.4$, calculate the forecast for the next four quarters (Year 3).

- *Initial Level (L_4): 120*
- *Initial Trend (T_4): 5*
- *Initial Seasonal Components: $S_1 = -15$, $S_2 = 5$, $S_3 = 30$, $S_4 = -10$*

Solution: The formulas for the Holt-Winters Additive Seasonal Model are applied iteratively for each quarter. Given $s = 4$, $\alpha = 0.2$, $\beta = 0.3$, and $\gamma = 0.4$.

- **For Quarter 5 (Year 2, Q1):**

$$\begin{aligned}
y_5 &= 110 \\
L_5 &= 0.2(110 - S_1) + 0.8(L_4 + T_4) \\
&= 0.2(110 - (-15)) + 0.8(120 + 5) \\
&= 0.2(125) + 0.8(125) = 125 \\
T_5 &= 0.3(L_5 - L_4) + 0.7T_4 \\
&= 0.3(125 - 120) + 0.7(5) = 0.3(5) + 3.5 = 5
\end{aligned}$$

$$\begin{aligned}
S_5 &= 0.4(110 - L_4 - T_4) + 0.6S_1 \\
&= 0.4(110 - 120 - 5) + 0.6(-15) = 0.4(-15) - 9 = -15 \\
S_5 &= 0.4(110 - L_4 - T_4) + 0.6S_1 \\
&= 0.4(110 - 120 - 5) + 0.6(-15) = 0.4(-15) - 9 = -15 \\
S_5 &= 0.4(110 - L_4 - T_4) + 0.6S_1 \\
&= 0.4(110 - 120 - 5) + 0.6(-15) = 0.4(-15) - 9 = -15 \\
\widehat{y}_6 &= L_5 + 1T_5 + S_{5+1-4} = 125 + 5 + S_2 = 130 + 5 = 135
\end{aligned}$$

- **For Quarter 6 (Year 2, Q2):** $y_6 = 135$, $L_6 = 0.2(135 - S_2) + 0.8(L_5 + T_5) = 0.2(135 - 5) + 0.8(125 + 5) = 0.2(130) + 0.8(130) = 130$, $T_6 = 0.3(L_6 - L_5) + 0.7T_5 = 0.3(130 - 125) + 0.7(5) = 0.3(5) + 3.5 = 5$, $S_6 = 0.4(135 - L_5 - T_5) + 0.6S_2 = 0.4(135 - 125 - 5) + 0.6(5) = 0.4(5) + 3 = 5$, $\widehat{y}_7 = L_6 + 1T_6 + S_{6+1-4} = 130 + 5 + S_3 = 135 + 30 = 165$

DRAFT

- **For Quarter 7 (Year 2, Q3):** $y_7 = 160$, $L_7 = 0.2(160-S_3)+0.8(L_6+T_6) = 0.2(160-30)+0.8(130+5) = 0.2(130)+0.8(135) = 26+108 = 134$, $T_7 = 0.3(L_7-L_6)+0.7T_6 = 0.3(134-130)+0.7(5) = 0.3(4)+3.5 = 4.7$, $S_7 = 0.4(160-L_6-T_6)+0.6S_3 = 0.4(160-130-5)+0.6(30) = 0.4(25)+18 = 28$, $\widehat{y}_8 = L_7+1T_7+S_{7+1-4} = 134+4.7+S_4 = 138.7-10 = 128.7$

- **For Quarter 8 (Year 2, Q4):** $y_8 = 125$, $L_8 = 0.2(125-S_4)+0.8(L_7+T_7) = 0.2(125-(-10))+0.8(134+4.7) = 0.2(135)+0.8(138.7) = 27+110.96 = 137.96$, $T_8 = 0.3(L_8-L_7)+0.7T_7 = 0.3(137.96-134)+0.7(4.7) = 0.3(3.96)+3.29 = 4.478$, $S_8 = 0.4(125-L_7-T_7)+0.6S_4 = 0.4(125-134-4.7)+0.6(-10) = 0.4(-13.7)-6 = -11.48$

Forecasting for Year 3: Using the last calculated values ($L_8 = 137.96$, $T_8 = 4.478$) and the updated seasonal components ($S_5 = -15, S_6 = 5, S_7 = 28, S_8 = -11.48$), we forecast the next four quarters.

- **Quarter 9 ($h = 1$):** $\widehat{y}_9 = L_8 + 1T_8 + S_{8+1-4} = L_8 + T_8 + S_5 = 137.96 + 4.478 + (-15) = 127.438$

- **Quarter 10 ($h = 2$):** $\widehat{y}_{10} = L_8 + 2T_8 + S_{8+2-4} = L_8 + 2T_8 + S_6 = 137.96 + 2(4.478) + 5 = 151.916$

- **Quarter 11 ($h = 3$):** $\widehat{y}_{11} = L_8 + 3T_8 + S_{8+3-4} = L_8 + 3T_8 + S_7 = 137.96 + 3(4.478) + 28 = 179.394$

- **Quarter 12 ($h = 4$):** $\widehat{y}_{12} = L_8 + 4T_8 + S_{8+4-4} = L_8 + 4T_8 + S_8 = 137.96 + 4(4.478) + (-11.48) = 144.392$

Advantages of Holt-Winters Seasonal Model

- **Comprehensive:** It can handle data with a combination of level, trend, and seasonality, making it a versatile tool for many real-world datasets.

- **Highly accurate for seasonal data:** When a time series has a stable seasonal pattern, this model often produces very accurate forecasts.

- **Two forms (additive and multiplicative):** The flexibility to choose between additive and multiplicative seasonality allows it to be applied to a wider range of data types.

Disadvantages of Holt-Winters Seasonal Model

- **Increased complexity:** The model is more complex to set up and interpret due to the three smoothing constants and the choice between additive and multiplicative seasonality.

DRAFT

- **Requires more data:** To get reliable initial values for the seasonal components, the model requires at least one full seasonal cycle of historical data. For monthly data, this means at least 12 data points.
- **Parameter initialization:** The initial values for the level, trend, and seasonal components can be difficult to determine without sufficient data or an established method.

Python Code for Holt-Winters Seasonal Model

```
from statsmodels.tsa.holtwinters import
  ExponentialSmoothing

# Fit Holt-Winters model with additive seasonality
model_hw_add = ExponentialSmoothing(data['Passengers'],
  seasonal='add', seasonal_periods=12).fit()

# Forecast future values
data['HW_Add'] = model_hw_add.fittedvalues

# Plot the original data and Holt-Winters model
plt.figure(figsize=(12, 6))
plt.plot(data['Passengers'], label='Original Data')
plt.plot(data['HW_Add'], label='Holt-Winters Additive
  Model', color='blue')
plt.title('Holt-Winters Seasonal Model (Additive)')
plt.xlabel('Time')
plt.ylabel('Number of Passengers')
plt.legend()
plt.show()
```

12.19 The ETS Framework

The ETS (Error, Trend, Seasonality) framework is a systematic approach to classifying and selecting exponential smoothing models. Each model is identified by a three-letter code, representing the type of component for its error, trend, and seasonality, respectively. This framework provides a structured and comprehensive way to define all possible exponential smoothing models.

12.19.1 Components of the ETS Framework

- **Error (E):** This component describes the nature of the residual errors.
  - **Additive (A):** The error variance is constant and does not depend on the level of the series.

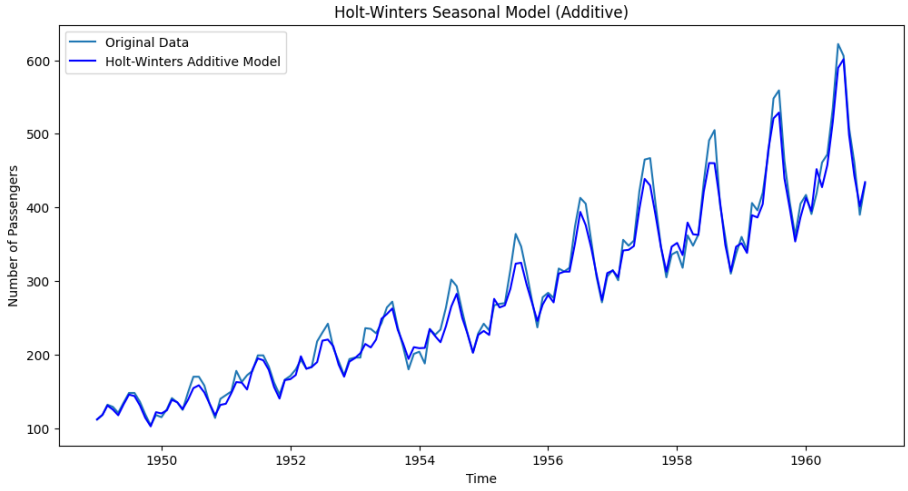


Figure 12.22: Holt-Winters Seasonal Model (Additive) of Air Passengers Data

- **Multiplicative (M):** The error variance is proportional to the level of the series. This is suitable for data where the magnitude of the fluctuations increases as the series' level increases.

- **Trend (T):** This component describes the long-term direction of the data.
  - **None (N):** The time series does not have a trend.
  - **Additive (A):** The trend increases or decreases by a constant amount.
  - **Multiplicative (M):** The trend increases or decreases by a constant percentage.
  - **Additive Damped (A_d):** An additive trend that is gradually reduced or "dampened" over time.
  - **Multiplicative Damped (M_d):** A multiplicative trend that is gradually dampened over time.

- **Seasonality (S):** This component captures repeating patterns over a fixed period.
  - **None (N):** The time series has no seasonal pattern.
  - **Additive (A):** The seasonal effect is constant over time.
  - **Multiplicative (M):** The seasonal effect changes in proportion to the level of the series.

DRAFT

12.19.2 Common ETS Models

The combination of these components results in various models. Some of the most common models are:

- **ETS(A,N,N):** This is the notation for **Simple Exponential Smoothing (SES)**. It assumes additive errors, no trend, and no seasonality. The forecast is simply the smoothed average of past data.
- **ETS(A,A,N):** This is the notation for **Holt's Linear Trend Model**. It includes an additive trend but no seasonality. The forecast is a combination of the smoothed level and a linear trend.
- **ETS(A,A,A):** This is the notation for the **Holt-Winters Additive Seasonal Model**. It assumes an additive trend and additive seasonality. This model is suitable for data where the seasonal fluctuations are of a constant magnitude.
- **ETS(A,A,M):** This is the notation for the **Holt-Winters Multiplicative Seasonal Model**. It assumes an additive trend and multiplicative seasonality. This model is ideal for data where the seasonal fluctuations increase or decrease with the level of the series.

12.19.3 Mathematical Formulation

The general form of an ETS model can be expressed in state space form, but a simpler, component-based view provides a clear understanding of the update equations for the level (L_t), trend (T_t), and seasonality (S_t).

Model: ETS(A,A,M)

For a model with additive errors, additive trend, and multiplicative seasonality, the component-wise equations are:

- **Level:** $L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1})$
- **Trend:** $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$
- **Seasonality:** $S_t = \gamma \frac{y_t}{L_t} + (1-\gamma)S_{t-s}$
- **Forecast:** $\widehat{y}_{t+h} = (L_t + hT_t)S_{t+h-s}$

where α, β, γ are the smoothing parameters and s is the seasonal cycle length.

DRAFT

Advantages of ETS

- **Comprehensive and Versatile:** The ETS framework provides a complete system for exponential smoothing models, allowing it to handle a wide variety of time series data. You can choose a model that fits your data's specific characteristics, whether it has no trend or seasonality (ETS(A,N,N)), a trend but no seasonality (ETS(A,A,N)), or both trend and seasonality (ETS(A,A,A) or ETS(A,A,M)).

- **Automatic Model Selection:** Statistical software packages can automatically select the best ETS model for a given dataset by minimizing information criteria like the **Akaike Information Criterion (AIC)**. This automates the process of choosing the right model, which was historically a manual and difficult task.

- **Effective for Short- to Medium-Term Forecasting:** ETS models are particularly effective for short- to medium-term forecasting. They are designed to adapt quickly to recent changes in the data, making them reliable for predicting the immediate future.

- **Easy to Understand and Interpret:** The models are based on intuitive components (level, trend, seasonality) that are easy to explain to non-technical audiences. The smoothing parameters directly relate to how much weight is given to new information.

*

Disadvantages of ETS

- **Limited to Specific Patterns:** The models are built on assumptions about the underlying data patterns (constant or linearly changing level, trend, and seasonality). They may not perform well on time series with non-linear trends, multiple seasonal cycles, or sudden, significant structural changes.

- **Parameter Sensitivity:** While the framework can automate parameter selection, the performance of the model is sensitive to the values of the smoothing parameters (α, β, γ). A poor choice of parameters can lead to inaccurate forecasts.

- **Challenges with Data Requirements:** Models with seasonality (e.g., ETS(A,A,A)) require at least one full seasonal cycle of data to be properly initialized. For data with long seasonal cycles (e.g., 12 months), this can be a significant limitation if you only have a small dataset.

- **Limited Handling of External Factors:** ETS models are univariate, meaning they only use a time series's past values to make pre-

DRAFT

dictions. They do not incorporate external variables (e.g., economic indicators, weather) that might influence the series, which can lead to less accurate forecasts in some cases.

12.20 Machine Learning Techniques for Time Series Analysis

Machine learning approaches to time series analysis have gained significant traction due to their ability to capture complex non-linear patterns and relationships in temporal data. Unlike traditional statistical methods, machine learning techniques can automatically learn features and patterns from data without requiring explicit specification of the underlying model structure.

12.20.1 ANN for Time Series Forecasting

An Artificial Neural Network (ANN) is a computational model inspired by the human brain's neural networks, used for time series forecasting. It provide a powerful framework for time series forecasting by learning complex non-linear mappings from historical data to future values. Unlike traditional statistical methods, ANNs can automatically discover patterns without requiring explicit model specification.

How ANNs Work for Time Series Forecasting

For time series forecasting, an Artificial Neural Network (ANN) takes a sequence of past data points as input to predict future values. The process generally involves:

1. Input Layer

The input layer receives a "window" of past data points. For example, to predict tomorrow's stock price, the input might be the prices from the last 30 days.

2. Hidden Layers

These layers consist of interconnected neurons that perform calculations on the input data. Each neuron applies a weight (w) and a bias (b) to its inputs (x) and passes the result through an activation function (f). This can be expressed mathematically as:

$$\text{Output} = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

These layers are where the network learns the underlying patterns and trends.

DRAFT

3. Output Layer

This layer produces the forecast. For a single-step forecast, it might have one neuron outputting the predicted value. For a multi-step forecast, it would have multiple neurons, one for each future time step being predicted.

Learning Process

The network learns by adjusting the weights and biases of its neurons to minimize the difference between its predictions and the actual observed values (the error). This process, called **backpropagation**, is repeated over many training epochs until the model achieves a satisfactory level of accuracy.

Types of ANNs for Time Series Forecasting

While a simple feedforward ANN can be used, more specialized architectures are often better suited for the sequential nature of time series data.

1. **Feedforward Neural Networks (FNNs):** The simplest form of ANN, where data flows in one direction from input to output. They are effective for simpler, stationary time series but may struggle with long-term dependencies.

2. **Recurrent Neural Networks (RNNs):** Unlike FNNs, RNNs have loops, allowing them to retain information from previous time steps. This "memory" makes them well-suited for sequential data. However, basic RNNs can suffer from the **vanishing gradient problem**, making it difficult to learn long-term dependencies.

3. **Long Short-Term Memory (LSTM) Networks:** An advanced type of RNN specifically designed to overcome the vanishing gradient problem. LSTMs use a system of "gates" (input, forget, and output gates) to regulate the flow of information, allowing them to selectively remember or forget past information. This makes them highly effective for time series with complex, long-range dependencies.

4. **Gated Recurrent Unit (GRU) Networks:** A simplified version of LSTMs that also addresses the vanishing gradient problem. GRUs have fewer gates (update and reset gates) and are computationally less complex, often providing comparable performance to LSTMs with faster training times.

12.20.2 Feedforward Neural Networks (FNNs)

Given a time series $\{y_1, y_2, \ldots, y_T\}$, the goal is to predict future values

$$\{y_{T+1}, y_{T+2}, \ldots, y_{T+h}\}$$

DRAFT

where h is the forecast horizon.

For ANN-based forecasting, we transform the time series into a supervised learning problem using a sliding window approach:

$$\begin{aligned} \text{Input:} \quad & \mathbf{X}_t = [y_{t-p+1}, y_{t-p+2}, \ldots, y_t] \in \mathbb{R}^p \\ \text{Target:} \quad & \text{target}_t = y_{t+1} \quad \text{(or multiple future values)} \end{aligned}$$

where p is the window size (number of lagged inputs).

A **Multi-Layer Perceptron (MLP)** is a specific type of FNN that consists of at least three layers: an input layer, one or more hidden layers, and an output layer. A typical MLP for time series forecasting consists of:

- **Input layer**: p neurons (lagged values)
- **Hidden layer(s)**: h neurons with activation functions
- **Output layer**: 1 or more neurons (forecast values)

Mathematical Formulation

For a single hidden layer MLP:

Hidden Layer:

$$z_j^{(1)} = \sum_{i=1}^{p} w_{ij}^{(1)} y_{t-p+i} + b_j^{(1)} \quad j = 1, 2, \ldots, h$$

$$a_j^{(1)} = f(z_j^{(1)}) = \frac{1}{1 + e^{-z_j^{(1)}}} \quad \text{(sigmoid activation)}$$

Output Layer:

$$z^{(2)} = \sum_{j=1}^{h} w_j^{(2)} a_j^{(1)} + b^{(2)}$$

$$\widehat{y}_{t+1} = z^{(2)} \quad \text{(linear output for regression)}$$

Matrix Form:

$$\begin{aligned} \mathbf{z}^{(1)} &= \mathbf{W}^{(1)} \mathbf{y}_t + \mathbf{b}^{(1)} \\ \mathbf{a}^{(1)} &= f(\mathbf{z}^{(1)}) \\ \widehat{y}_{t+1} &= \mathbf{w}^{(2)T} \mathbf{a}^{(1)} + b^{(2)} \end{aligned}$$

where:

- $\mathbf{W}^{(1)} \in \mathbb{R}^{h \times p}$: input-to-hidden weights
- $\mathbf{w}^{(2)} \in \mathbb{R}^h$: hidden-to-output weights
- $\mathbf{b}^{(1)} \in \mathbb{R}^h$, $b^{(2)} \in \mathbb{R}$: bias terms

DRAFT

Training Algorithm: Backpropagation

Loss Function

For regression problems, we typically use Mean Squared Error:

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Gradient Computation

Using the chain rule, gradients are computed as:

Output Layer Gradients:

$$\frac{\partial L}{\partial w_j^{(2)}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)a_j^{(1)}$$
$$\frac{\partial L}{\partial b^{(2)}} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)$$

Hidden Layer Gradients:

$$\delta_j^{(1)} = (\widehat{y}_i - y_i)w_j^{(2)} f'(z_j^{(1)})$$
$$\frac{\partial L}{\partial w_{ij}^{(1)}} = \frac{1}{n}\sum_{i=1}^{n}\delta_j^{(1)} x_i$$
$$\frac{\partial L}{\partial b_j^{(1)}} = \frac{1}{n}\sum_{i=1}^{n}\delta_j^{(1)}$$

Weight Update

Parameters are updated using gradient descent:

$$\theta_{new} = \theta_{old} - \alpha\frac{\partial L}{\partial \theta}$$

where α is the learning rate.

Problem 12.9. *Consider a simple stock price time series forecasting problem. Let's use the following daily stock prices over 10 days:*

Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Price ($)	*100.0*	*102.5*	*101.8*	*104.2*	*106.1*	*105.3*	*107.8*	*109.2*	*108.5*	*110.1*

Show the data processing steps and first hidden layer for time series forecasting.

DRAFT

Solution

We are given a simple stock price time series for 10 days. To prepare the data for time series forecasting using an artificial neural network, we follow these steps:

Data Preparation:

- We choose a window size (look-back period). Suppose we use a window of $p = 3$ days to predict the next day's price.
- The training samples are constructed as:

$$\begin{aligned} &\text{Input (Day 1–3)} \rightarrow \text{Output (Day 4)} \\ &(100.0, 102.5, 101.8) \rightarrow 104.2 \\ &(102.5, 101.8, 104.2) \rightarrow 106.1 \\ &(101.8, 104.2, 106.1) \rightarrow 105.3 \\ &\qquad\qquad \vdots \\ &(107.8, 109.2, 108.5) \rightarrow 110.1 \end{aligned}$$

- Thus, we have 7 training samples.

Step 1: Normalization Stock prices are usually normalized (e.g., using Min–Max scaling or Standardization) before being fed into the network. For illustration, let us consider Min–Max scaling. Normalize the data to $[0, 1]$ range:

$$y_{norm} = \frac{y - y_{min}}{y_{max} - y_{min}}$$

For our data: $y_{min} = 100.0$, $y_{max} = 110.1$

Normalized prices: $[0.000, 0.248, 0.178, 0.416, 0.604, 0.525, 0.772, 0.911, 0.842, 1.000]$

Step 2: Create Training Samples

Using window size $p = 3$ (using 3 previous days to predict next day):

Sample	y_{t-2}	y_{t-1}	y_t	Target y_{t+1}
1	0.000	0.248	0.178	0.416
2	0.248	0.178	0.416	0.604
3	0.178	0.416	0.604	0.525
4	0.416	0.604	0.525	0.772
5	0.604	0.525	0.772	0.911
6	0.525	0.772	0.911	0.842
7	0.772	0.911	0.842	1.000

DRAFT

Network Configuration

- Input neurons: 3 (window size). Each input is a 3-dimensional vector representing the past 3 days' prices.
- Hidden neurons: 5
- Output neurons: 1
- Activation function: Sigmoid for hidden layer, Linear for output
- Learning rate: $\alpha = 0.1$

Forward Pass Calculation

For the first training sample with input sequence $[y_{t-2}, y_{t-1}, y_t] = [0.000, 0.248, 0.178]$ and target $y_{t+1} = 0.416$:

Step 1: Hidden Layer Computation

Assume initial weights and biases:

$$\mathbf{W}^{(1)} = \begin{bmatrix} 0.5 & -0.3 & 0.8 \\ 0.2 & 0.7 & -0.4 \\ -0.6 & 0.1 & 0.9 \\ 0.3 & -0.2 & 0.6 \\ 0.4 & 0.8 & -0.1 \end{bmatrix}, \quad \mathbf{b}^{(1)} = \begin{bmatrix} 0.1 \\ -0.2 \\ 0.3 \\ 0.0 \\ -0.1 \end{bmatrix}$$

Computing each hidden neuron:

Hidden Neuron 1:

$$\begin{aligned} z_1^{(1)} &= w_{11}^{(1)} y_{t-2} + w_{12}^{(1)} y_{t-1} + w_{13}^{(1)} y_t + b_1^{(1)} \\ &= 0.5 \times 0.000 + (-0.3) \times 0.248 + 0.8 \times 0.178 + 0.1 \\ &= 0.000 - 0.074 + 0.142 + 0.1 = 0.168 \end{aligned}$$

$$a_1^{(1)} = \frac{1}{1 + e^{-0.168}} = \frac{1}{1 + 0.845} = 0.542$$

Hidden Neuron 2:

$$\begin{aligned} z_2^{(1)} &= w_{21}^{(1)} y_{t-2} + w_{22}^{(1)} y_{t-1} + w_{23}^{(1)} y_t + b_2^{(1)} \\ &= 0.2 \times 0.000 + 0.7 \times 0.248 + (-0.4) \times 0.178 + (-0.2) \\ &= 0.000 + 0.174 - 0.071 - 0.2 = -0.097 \end{aligned}$$

$$a_2^{(1)} = \frac{1}{1 + e^{-(-0.097)}} = \frac{1}{1 + e^{0.097}} = \frac{1}{1 + 1.102} = 0.476$$

DRAFT

Hidden Neuron 3:

$$\begin{aligned} z_3^{(1)} &= w_{31}^{(1)} y_{t-2} + w_{32}^{(1)} y_{t-1} + w_{33}^{(1)} y_t + b_3^{(1)} \\ &= (-0.6) \times 0.000 + 0.1 \times 0.248 + 0.9 \times 0.178 + 0.3 \\ &= 0.000 + 0.025 + 0.160 + 0.3 = 0.485 \\ a_3^{(1)} &= \frac{1}{1 + e^{-0.485}} = \frac{1}{1 + 0.616} = 0.619 \end{aligned}$$

Hidden Neuron 4:

$$\begin{aligned} z_4^{(1)} &= w_{41}^{(1)} y_{t-2} + w_{42}^{(1)} y_{t-1} + w_{43}^{(1)} y_t + b_4^{(1)} \\ &= 0.3 \times 0.000 + (-0.2) \times 0.248 + 0.6 \times 0.178 + 0.0 \\ &= 0.000 - 0.050 + 0.107 + 0.0 = 0.057 \\ a_4^{(1)} &= \frac{1}{1 + e^{-0.057}} = \frac{1}{1 + 0.945} = 0.514 \end{aligned}$$

Hidden Neuron 5:

$$\begin{aligned} z_5^{(1)} &= w_{51}^{(1)} y_{t-2} + w_{52}^{(1)} y_{t-1} + w_{53}^{(1)} y_t + b_5^{(1)} \\ &= 0.4 \times 0.000 + 0.8 \times 0.248 + (-0.1) \times 0.178 + (-0.1) \\ &= 0.000 + 0.198 - 0.018 - 0.1 = 0.080 \\ a_5^{(1)} &= \frac{1}{1 + e^{-0.080}} = \frac{1}{1 + 0.923} = 0.520 \end{aligned}$$

Therefore:

$$\mathbf{a}^{(1)} = [0.542, 0.476, 0.619, 0.514, 0.520]$$

Step 2: Output Layer Computation
Assume output layer weights and bias:

$$\mathbf{w}^{(2)} = [0.6, -0.4, 0.8, 0.2, -0.3], \quad b^{(2)} = 0.05$$

$$\begin{aligned} \widehat{y}_{t+1} &= \sum_{j=1}^{5} w_j^{(2)} a_j^{(1)} + b^{(2)} \\ &= 0.6 \times 0.542 + (-0.4) \times 0.476 + 0.8 \times 0.619 \\ &\quad + 0.2 \times 0.514 + (-0.3) \times 0.520 + 0.05 \\ &= 0.325 - 0.190 + 0.495 + 0.103 - 0.156 + 0.05 \\ &= 0.627 \end{aligned}$$

Step 3: Error Computation

$$\text{Error} = \widehat{y}_{t+1} - y_{t+1} = 0.627 - 0.416 = 0.211$$

DRAFT

Back-transformation of Forecast: For Min–Max scaling to $[0, 1]$ with $y_{norm} = (y - y_{\min})/(y_{\max} - y_{\min})$, invert via:

$$\widehat{y} = \widehat{y}_{norm} \times (y_{\max} - y_{\min}) + y_{\min}$$

With $y_{\min} = 100.0$, $y_{\max} = 110.1$ and $\widehat{y}_{t+1} = 0.627$:

$$\begin{aligned}\widehat{y}_{t+1} &= 0.627 \times (110.1 - 100.0) + 100.0 \\ &= 0.627 \times 10.1 + 100.0 = 6.33 + 100.0 = 106.33\end{aligned}$$

So the original-scale error is $106.33 - 104.2 = 2.13$.

This comprehensive framework demonstrates how ANNs can be effectively applied to time series forecasting problems, providing both theoretical foundations and practical implementation guidance.

Implementation Considerations

- **Data Splitting**: Use time-based splits (not random) to avoid data leakage
- **Stationarity**: Consider differencing or detrending non-stationary series
- **Feature Engineering**: Include additional features like seasonality indicators, external variables
- **Hyperparameter Tuning**: Optimize window size, hidden units, learning rate using validation data
- **Early Stopping**: Monitor validation loss to prevent overfitting

12.20.3 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks are a class of neural networks specifically designed to handle sequential data, such as time series, by maintaining an internal "memory" through hidden states. Unlike traditional feedforward networks, RNNs process a sequence one element at a time, using information from previous steps to inform the current one.

Vanilla RNN

The basic or "vanilla" RNN processes a sequence $\{x_1, x_2, \ldots, x_T\}$ through the following core equations. These equations describe how the network updates its hidden state and produces an output at each time step t.

$$\begin{aligned}h_t &= \tanh(W_h h_{t-1} + W_x x_t + b_h) \\ y_t &= W_y h_t + b_y\end{aligned}$$

Here's a breakdown of the components:

DRAFT

- $h_t \in \mathbb{R}^d$ is the **hidden state** at time t. This is the network's memory, holding information from all previous inputs.
- $x_t \in \mathbb{R}^n$ is the **input vector** at time t. For a time series like stock prices, n would be 1 if we only use the price.
- $y_t \in \mathbb{R}^m$ is the **output vector** at time t. This could be the forecasted value for the next time step.
- $W_h \in \mathbb{R}^{d\times d}$, $W_x \in \mathbb{R}^{d\times n}$, $W_y \in \mathbb{R}^{m\times d}$ are the **weight matrices** that the network learns during training.
- $b_h \in \mathbb{R}^d$, $b_y \in \mathbb{R}^m$ are **bias vectors** that are also learned during training.

Problem 12.10. ***Example: Stock Price Forecasting***

To demonstrate a step-by-step calculation, we will use a simple, single-input RNN to forecast a stock price. Let's assume we have a trained model with predefined, hypothetical parameters. For simplicity, we will use a hidden state of size $d = 1$ and an output of size $m = 1$. The daily stock prices for the last 5 days are:

Day	*1*	*2*	*3*	*4*	*5*
Price ($)	*100.0*	*102.5*	*101.8*	*104.2*	*106.1*

Solution

We are given a simple stock price time series for 5 days. To prepare the data for time series forecasting using a vanilla RNN, we follow these steps:
Data Preprocessing: In practice, RNNs work better with normalized inputs. We normalize the prices using:

$$x_t^{(normalized)} = \frac{\text{Price}_t - \mu}{\sigma}$$

where $\mu = 102.92$ (mean) and $\sigma = 2.26$ (standard deviation). The normalized inputs become:

Day	1	2	3	4	5
Normalized Input	-1.29	-0.19	-0.50	0.57	1.41

Model Parameters: For this example, we assume the RNN has been trained and has the following realistic parameters:

- $W_x = 0.5$
- $W_h = 0.3$

DRAFT

- $W_y = 2.0$
- $b_h = 0.1$
- $b_y = 0.8$

We also assume the initial hidden state $h_0 = 0$.

Step-by-Step Calculation:

Day 1 ($t = 1$):

$$\begin{aligned} x_1 &= -1.29 \\ h_1 &= \tanh(W_h h_0 + W_x x_1 + b_h) \\ &= \tanh(0.3 \cdot 0 + 0.5 \cdot (-1.29) + 0.1) \\ &= \tanh(-0.645 + 0.1) = \tanh(-0.545) = -0.497 \end{aligned}$$

Day 2 ($t = 2$):

$$\begin{aligned} x_2 &= -0.19 \\ h_2 &= \tanh(W_h h_1 + W_x x_2 + b_h) \\ &= \tanh(0.3 \cdot (-0.497) + 0.5 \cdot (-0.19) + 0.1) \\ &= \tanh(-0.149 - 0.095 + 0.1) = \tanh(-0.144) = -0.143 \end{aligned}$$

Day 3 ($t = 3$):

$$\begin{aligned} x_3 &= -0.50 \\ h_3 &= \tanh(W_h h_2 + W_x x_3 + b_h) \\ &= \tanh(0.3 \cdot (-0.143) + 0.5 \cdot (-0.50) + 0.1) \\ &= \tanh(-0.043 - 0.250 + 0.1) = \tanh(-0.193) = -0.190 \end{aligned}$$

Day 4 ($t = 4$):

$$\begin{aligned} x_4 &= 0.57 \\ h_4 &= \tanh(W_h h_3 + W_x x_4 + b_h) \\ &= \tanh(0.3 \cdot (-0.190) + 0.5 \cdot 0.57 + 0.1) \\ &= \tanh(-0.057 + 0.285 + 0.1) = \tanh(0.328) = 0.316 \end{aligned}$$

Day 5 ($t = 5$):

$$\begin{aligned} x_5 &= 1.41 \\ h_5 &= \tanh(W_h h_4 + W_x x_5 + b_h) \\ &= \tanh(0.3 \cdot 0.316 + 0.5 \cdot 1.41 + 0.1) \\ &= \tanh(0.095 + 0.705 + 0.1) = \tanh(0.900) = 0.716 \end{aligned}$$

DRAFT

Forecast for Day 6: The forecast is the output y_5 using the final hidden state h_5.

$$\begin{aligned} y_5 &= W_y h_5 + b_y \\ &= 2.0 \cdot 0.716 + 0.8 = 1.432 + 0.8 = 2.232 \end{aligned}$$

Converting Back to Price: To get the actual predicted price, we denormalize the output:

$$\begin{aligned} \text{Predicted Price} &= y_5 \times \sigma + \mu \\ &= 2.232 \times 2.26 + 102.92 \\ &= 5.04 + 102.92 = 107.96 \text{ dollars} \end{aligned}$$

This prediction suggests the stock price will increase to approximately \$107.96 on Day 6, which follows the upward trend observed in the recent data. □

12.21 LSTM for Time Series Forecasting

Long Short-Term Memory (LSTM) is a specialized type of recurrent neural network (RNN) designed to effectively learn and retain long-term dependencies in sequential data. Traditional RNNs struggle with vanishing and exploding gradient problems, which hinder their ability to capture relationships over extended sequences. LSTM overcomes this limitation by introducing a unique memory cell structure that selectively retains or forgets information as needed. Time series forecasting is one of the most successful applications of LSTM networks, where the ability to capture long-term dependencies becomes crucial for predicting future values based on historical patterns. Unlike traditional statistical methods, LSTMs can automatically learn complex nonlinear relationships and temporal dependencies in sequential data.

12.21.1 LSTM Architecture for Time Series

For time series forecasting, an LSTM processes a sequence of historical observations $\{x_1, x_2, \ldots, x_T\}$ to predict future values $\{x_{T+1}, x_{T+2}, \ldots, x_{T+h}\}$, where h is the forecasting horizon. The architecture of LSTM (See the Figure 12.23) revolves around its three main gates—forget, input, and output gates—that work together to manage the flow of information in the network. The LSTM maintains two types of memory:

- **Cell State** (C_t): Long-term memory that flows through the network with minimal linear interactions
- **Hidden State** (h_t): Short-term memory that serves as the output at each time step

DRAFT

The LSTM equations for time series forecasting are:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(candidate values)} \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(cell state update)} \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \\
h_t &= o_t \odot \tanh(C_t) \quad \text{(hidden state)}
\end{aligned}
$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, and $[\cdot, \cdot]$ represents concatenation.

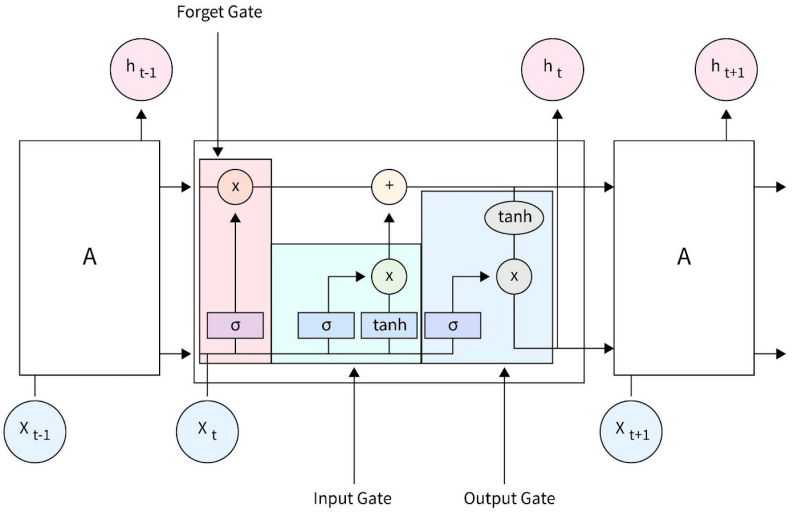


Figure 12.23: The architecture of Long Short-Term Memory (LSTM) revolves around its three main gates—forget, input, and output gates.

The final prediction is typically generated using a fully connected layer:

$$\widehat{y}_t = W_y h_t + b_y \tag{12.5}$$

12.21.2 Practical Example: Stock Price Forecasting

To illustrate the LSTM forecasting process, consider a simple stock price prediction problem using the following daily prices over 5 days:

Day	1	2	3	4	5
Price ($)	100.0	102.5	101.8	104.2	106.1

Step 1: Data Preprocessing

Time series data often requires normalization to improve training stability and convergence. We apply z-score normalization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{12.6}$$

where $\mu = 102.92$ and $\sigma = 2.26$ for our dataset.
The normalized sequence becomes:

$$\mathbf{x}_{\text{norm}} = [-1.29, -0.19, -0.50, 0.57, 1.41]$$

Step 2: Model Configuration

For demonstration purposes, we use a simplified LSTM with:

- Input dimension: $d_i = 1$ (single price value)
- Hidden dimension: $d_h = 2$
- Concatenated input dimension: $d_{concat} = d_h + d_i = 3$
- Sequence length: $L = 3$ (using last 3 days to predict next day)

The weight matrices are initialized with proper dimensions ($d_h \times d_{concat}$):

$$W_f = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ -0.1 & 0.4 & 0.2 \end{bmatrix}, \quad b_f = \begin{bmatrix} 0.1 \\ 0.0 \end{bmatrix}$$

$$W_i = \begin{bmatrix} 0.3 & -0.2 & 0.1 \\ 0.1 & 0.3 & -0.1 \end{bmatrix}, \quad b_i = \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$$

$$W_C = \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.2 \end{bmatrix}, \quad b_C = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.2 & -0.3 & 0.1 \end{bmatrix}, \quad b_o = \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 1.0 & 0.5 \end{bmatrix}, \quad b_y = 0.0$$

DRAFT

Step 3: Forward Pass Computation

We use the sequence $[x_3, x_4, x_5] = [-0.50, 0.57, 1.41]$ to predict x_6.

Time Step $t = 1$ (Day 3, $x_3 = -0.50$):
Initialize: $h_0 = [0, 0]^T$, $C_0 = [0, 0]^T$
Input concatenation: $[h_0; x_3] = [0, 0, -0.50]^T$
Forget Gate:

$$W_f \cdot [0, 0, -0.50]^T = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ -0.1 & 0.4 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -0.50 \end{bmatrix} = \begin{bmatrix} -0.15 \\ -0.10 \end{bmatrix}$$

$$f_1 = \sigma\left(\begin{bmatrix} -0.15 \\ -0.10 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.0 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} -0.05 \\ -0.10 \end{bmatrix}\right) = \begin{bmatrix} 0.488 \\ 0.475 \end{bmatrix}$$

Input Gate:

$$W_i \cdot [0, 0, -0.50]^T = \begin{bmatrix} 0.3 & -0.2 & 0.1 \\ 0.1 & 0.3 & -0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -0.50 \end{bmatrix} = \begin{bmatrix} -0.05 \\ 0.05 \end{bmatrix}$$

$$i_1 = \sigma\left(\begin{bmatrix} -0.05 \\ 0.05 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} -0.05 \\ 0.15 \end{bmatrix}\right) = \begin{bmatrix} 0.488 \\ 0.537 \end{bmatrix}$$

Candidate Values:

$$W_C \cdot [0, 0, -0.50]^T = \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -0.50 \end{bmatrix} = \begin{bmatrix} 0.05 \\ -0.10 \end{bmatrix}$$

$$\tilde{C}_1 = \tanh\left(\begin{bmatrix} 0.05 \\ -0.10 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}\right) = \tanh\left(\begin{bmatrix} 0.05 \\ -0.10 \end{bmatrix}\right) = \begin{bmatrix} 0.050 \\ -0.100 \end{bmatrix}$$

Cell State Update:

$$\begin{aligned} C_1 &= f_1 \odot C_0 + i_1 \odot \tilde{C}_1 \\ &= \begin{bmatrix} 0.488 \\ 0.475 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.488 \\ 0.537 \end{bmatrix} \odot \begin{bmatrix} 0.050 \\ -0.100 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.024 \\ -0.054 \end{bmatrix} = \begin{bmatrix} 0.024 \\ -0.054 \end{bmatrix} \end{aligned}$$

DRAFT

Output Gate and Hidden State:

$$W_o \cdot [0, 0, -0.50]^T = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.2 & -0.3 & 0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -0.50 \end{bmatrix} = \begin{bmatrix} -0.10 \\ -0.05 \end{bmatrix}$$

$$o_1 = \sigma\left(\begin{bmatrix} -0.10 \\ -0.05 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} -0.10 \\ 0.05 \end{bmatrix}\right) = \begin{bmatrix} 0.475 \\ 0.513 \end{bmatrix}$$

$$h_1 = o_1 \odot \tanh(C_1) = \begin{bmatrix} 0.475 \\ 0.513 \end{bmatrix} \odot \begin{bmatrix} 0.024 \\ -0.054 \end{bmatrix} = \begin{bmatrix} 0.011 \\ -0.028 \end{bmatrix}$$

Time Step $t = 2$ (Day 4, $x_4 = 0.57$):
Input concatenation: $[h_1; x_4] = [0.011, -0.028, 0.57]^T$
Forget Gate:

$$W_f \cdot [0.011, -0.028, 0.57]^T = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ -0.1 & 0.4 & 0.2 \end{bmatrix} \begin{bmatrix} 0.011 \\ -0.028 \\ 0.57 \end{bmatrix}$$

$$= \begin{bmatrix} 0.002 - 0.003 + 0.171 \\ -0.001 - 0.011 + 0.114 \end{bmatrix} = \begin{bmatrix} 0.170 \\ 0.102 \end{bmatrix}$$

$$f_2 = \sigma\left(\begin{bmatrix} 0.170 \\ 0.102 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.0 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 0.270 \\ 0.102 \end{bmatrix}\right) = \begin{bmatrix} 0.567 \\ 0.525 \end{bmatrix}$$

Input Gate:

$$W_i \cdot [0.011, -0.028, 0.57]^T = \begin{bmatrix} 0.3 & -0.2 & 0.1 \\ 0.1 & 0.3 & -0.1 \end{bmatrix} \begin{bmatrix} 0.011 \\ -0.028 \\ 0.57 \end{bmatrix}$$

$$= \begin{bmatrix} 0.003 + 0.006 + 0.057 \\ 0.001 - 0.008 - 0.057 \end{bmatrix} = \begin{bmatrix} 0.066 \\ -0.064 \end{bmatrix}$$

$$i_2 = \sigma\left(\begin{bmatrix} 0.066 \\ -0.064 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 0.066 \\ 0.036 \end{bmatrix}\right) = \begin{bmatrix} 0.517 \\ 0.509 \end{bmatrix}$$

DRAFT

Candidate Values:

$$W_C \cdot [0.011, -0.028, 0.57]^T = \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0.011 \\ -0.028 \\ 0.57 \end{bmatrix}$$

$$= \begin{bmatrix} 0.001 - 0.006 - 0.057 \\ 0.003 + 0.003 + 0.114 \end{bmatrix} = \begin{bmatrix} -0.062 \\ 0.120 \end{bmatrix}$$

$$\tilde{C}_2 = \tanh\left(\begin{bmatrix} -0.062 \\ 0.120 \end{bmatrix} \right) = \begin{bmatrix} -0.062 \\ 0.119 \end{bmatrix}$$

Cell State Update:

$$C_2 = f_2 \odot C_1 + i_2 \odot \tilde{C}_2$$

$$= \begin{bmatrix} 0.567 \\ 0.525 \end{bmatrix} \odot \begin{bmatrix} 0.024 \\ -0.054 \end{bmatrix} + \begin{bmatrix} 0.517 \\ 0.509 \end{bmatrix} \odot \begin{bmatrix} -0.062 \\ 0.119 \end{bmatrix}$$

$$= \begin{bmatrix} 0.014 \\ -0.028 \end{bmatrix} + \begin{bmatrix} -0.032 \\ 0.061 \end{bmatrix} = \begin{bmatrix} -0.018 \\ 0.033 \end{bmatrix}$$

Output Gate and Hidden State:

$$W_o \cdot [0.011, -0.028, 0.57]^T = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.2 & -0.3 & 0.1 \end{bmatrix} \begin{bmatrix} 0.011 \\ -0.028 \\ 0.57 \end{bmatrix}$$

$$= \begin{bmatrix} 0.004 - 0.003 + 0.114 \\ 0.002 + 0.008 + 0.057 \end{bmatrix} = \begin{bmatrix} 0.115 \\ 0.067 \end{bmatrix}$$

$$o_2 = \sigma\left(\begin{bmatrix} 0.115 \\ 0.067 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} \right) = \sigma\left(\begin{bmatrix} 0.115 \\ 0.167 \end{bmatrix} \right) = \begin{bmatrix} 0.529 \\ 0.542 \end{bmatrix}$$

$$h_2 = o_2 \odot \tanh(C_2) = \begin{bmatrix} 0.529 \\ 0.542 \end{bmatrix} \odot \begin{bmatrix} -0.018 \\ 0.033 \end{bmatrix} = \begin{bmatrix} -0.010 \\ 0.018 \end{bmatrix}$$

Time Step $t = 3$ (Day 5, $x_5 = 1.41$):
Input concatenation: $[h_2; x_5] = [-0.010, 0.018, 1.41]^T$

DRAFT

Forget Gate:

$$W_f \cdot [-0.010, 0.018, 1.41]^T = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ -0.1 & 0.4 & 0.2 \end{bmatrix} \begin{bmatrix} -0.010 \\ 0.018 \\ 1.41 \end{bmatrix}$$

$$= \begin{bmatrix} -0.002 + 0.002 + 0.423 \\ 0.001 + 0.007 + 0.282 \end{bmatrix} = \begin{bmatrix} 0.423 \\ 0.290 \end{bmatrix}$$

$$f_3 = \sigma\left(\begin{bmatrix} 0.423 \\ 0.290 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.0 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 0.523 \\ 0.290 \end{bmatrix}\right) = \begin{bmatrix} 0.628 \\ 0.572 \end{bmatrix}$$

Input Gate:

$$W_i \cdot [-0.010, 0.018, 1.41]^T = \begin{bmatrix} 0.3 & -0.2 & 0.1 \\ 0.1 & 0.3 & -0.1 \end{bmatrix} \begin{bmatrix} -0.010 \\ 0.018 \\ 1.41 \end{bmatrix}$$

$$= \begin{bmatrix} -0.003 - 0.004 + 0.141 \\ -0.001 + 0.005 - 0.141 \end{bmatrix} = \begin{bmatrix} 0.134 \\ -0.137 \end{bmatrix}$$

$$i_3 = \sigma\left(\begin{bmatrix} 0.134 \\ -0.137 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 0.134 \\ -0.037 \end{bmatrix}\right) = \begin{bmatrix} 0.533 \\ 0.491 \end{bmatrix}$$

Candidate Values:

$$W_C \cdot [-0.010, 0.018, 1.41]^T = \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} -0.010 \\ 0.018 \\ 1.41 \end{bmatrix}$$

$$= \begin{bmatrix} -0.001 + 0.004 - 0.141 \\ -0.003 - 0.002 + 0.282 \end{bmatrix} = \begin{bmatrix} -0.138 \\ 0.277 \end{bmatrix}$$

$$\tilde{C}_3 = \tanh\left(\begin{bmatrix} -0.138 \\ 0.277 \end{bmatrix}\right) = \begin{bmatrix} -0.137 \\ 0.271 \end{bmatrix}$$

DRAFT

Cell State Update:

$$\begin{aligned} C_3 &= f_3 \odot C_2 + i_3 \odot \tilde{C}_3 \\ &= \begin{bmatrix} 0.628 \\ 0.572 \end{bmatrix} \odot \begin{bmatrix} -0.018 \\ 0.033 \end{bmatrix} + \begin{bmatrix} 0.533 \\ 0.491 \end{bmatrix} \odot \begin{bmatrix} -0.137 \\ 0.271 \end{bmatrix} \\ &= \begin{bmatrix} -0.011 \\ 0.019 \end{bmatrix} + \begin{bmatrix} -0.073 \\ 0.133 \end{bmatrix} = \begin{bmatrix} -0.084 \\ 0.152 \end{bmatrix} \end{aligned}$$

Output Gate and Hidden State:

$$W_o \cdot [-0.010, 0.018, 1.41]^T = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.2 & -0.3 & 0.1 \end{bmatrix} \begin{bmatrix} -0.010 \\ 0.018 \\ 1.41 \end{bmatrix}$$

$$= \begin{bmatrix} -0.004 + 0.002 + 0.282 \\ -0.002 - 0.005 + 0.141 \end{bmatrix} = \begin{bmatrix} 0.280 \\ 0.134 \end{bmatrix}$$

$$o_3 = \sigma\left(\begin{bmatrix} 0.280 \\ 0.134 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} \right) = \sigma\left(\begin{bmatrix} 0.280 \\ 0.234 \end{bmatrix} \right) = \begin{bmatrix} 0.570 \\ 0.558 \end{bmatrix}$$

$$h_3 = o_3 \odot \tanh(C_3) = \begin{bmatrix} 0.570 \\ 0.558 \end{bmatrix} \odot \begin{bmatrix} -0.084 \\ 0.151 \end{bmatrix} = \begin{bmatrix} -0.048 \\ 0.084 \end{bmatrix}$$

Step 4: Prediction Generation

The final prediction is generated using the linear output layer:

$$\begin{aligned} \widehat{y}_{\text{norm}} &= W_y h_3 + b_y \\ &= \begin{bmatrix} 1.0 & 0.5 \end{bmatrix} \begin{bmatrix} -0.048 \\ 0.084 \end{bmatrix} + 0.0 \\ &= 1.0 \times (-0.048) + 0.5 \times 0.084 + 0.0 \\ &= -0.048 + 0.042 + 0.0 = -0.006 \end{aligned}$$

Step 5: Denormalization

Converting back to the original scale:

$$\begin{aligned} \widehat{y} &= \widehat{y}_{\text{norm}} \times \sigma + \mu \\ &= -0.006 \times 2.26 + 102.92 \\ &= -0.014 + 102.92 = 102.91 \end{aligned}$$

DRAFT

Therefore, the predicted stock price for Day 6 is **$102.91**, which represents a significant decrease from the Day 5 price of $106.1, suggesting the model predicts a correction after the recent upward trend.

12.21.3 Key Advantages of LSTM for Time Series

1. **Long-term Memory**: LSTMs can retain information over long sequences, crucial for capturing seasonal patterns and long-range dependencies.
2. **Selective Memory**: The gating mechanism allows the model to forget irrelevant information and retain important patterns.
3. **Nonlinear Modeling**: Unlike linear time series methods, LSTMs can capture complex nonlinear relationships in the data.
4. **Multivariate Capability**: LSTMs can easily handle multiple input features, making them suitable for multivariate time series forecasting.
5. **End-to-End Learning**: The entire model can be trained end-to-end using backpropagation through time, automatically learning optimal feature representations.

12.21.4 Training Considerations

When training LSTMs for time series forecasting, several factors must be considered:

Loss Function: For regression tasks like stock price prediction, mean squared error (MSE) is commonly used:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 \tag{12.7}$$

Sequence Length: The choice of sequence length L is crucial. Too short sequences may not capture long-term patterns, while too long sequences can lead to vanishing gradients and increased computational cost.

Regularization: Techniques like dropout, weight decay, and early stopping help prevent overfitting, which is common in time series due to temporal correlations.

Data Splitting: Time series data requires careful splitting to maintain temporal order. Typically, earlier data is used for training and later data for validation and testing.

This example demonstrates the step-by-step computation of LSTM for time series forecasting, showing how the gating mechanisms work together to process sequential information and generate predictions based on historical patterns.

DRAFT

DRAFT

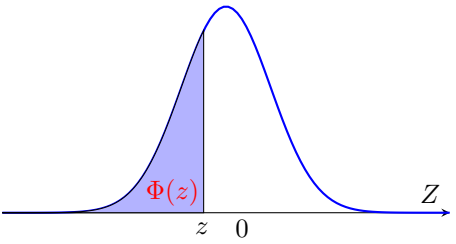
Chapter 13

Appendix

A.1: Table of 1000 random digits

01	32924	22324	18125	09077	26	96772	16443	39877	04653
02	54632	90374	94143	49295	27	52167	21038	14338	01395
03	88720	43035	97081	83373	28	69644	37198	00028	98195
04	21727	11904	41513	31653	29	71011	62004	81712	87536
05	80985	70799	57975	69282	30	31217	75877	85366	55500
06	40412	58826	94868	52632	31	64990	98735	02999	35521
07	43918	56807	75218	46077	32	48417	23569	59307	46550
08	26513	47480	77410	47741	33	07900	65059	48592	44087
09	18164	35784	44255	30124	34	74526	32601	24482	16981
10	39446	01375	75264	51173	35	51056	04402	58353	37332
11	16638	04680	98617	90298	36	39005	93458	63143	21817
12	16872	94749	44012	48884	37	67883	76343	78155	67733
13	65419	87092	78596	91512	38	06014	60999	87226	36071
14	05207	36702	56804	10498	39	93147	88766	04148	42471
15	78807	79243	13729	81222	40	01099	95731	47622	13294
16	69341	79028	64253	80447	41	89252	01201	58138	13809
17	41871	17566	61200	15994	42	41766	57239	50251	64675
18	25758	04625	43226	32986	43	92736	77800	81996	45646
19	06604	94486	40174	10742	44	45118	36600	68977	68831
20	82259	56512	48945	18183	45	73457	01579	00378	70197
21	07895	37090	50627	71320	46	49465	85251	42914	17277
22	59836	71148	42320	67816	47	15745	37285	23768	39302
23	57133	76610	89104	30481	48	28760	81331	78265	60690
24	76964	57126	87174	61025	49	82193	32787	70451	91141

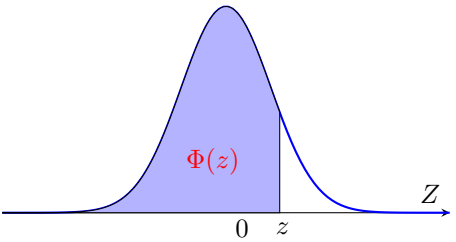
DRAFT



A.2: Cumulative Distribution Function $\Phi(z) = P(Z \leq z)$ for $z \leq 0$ of the Standard Normal Distribution

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
-3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

DRAFT



A.3: Cumulative Distribution Function $\Phi(z) = P(Z \leq z)$ for $z \geq 0$ of the Standard Normal Distribution

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.7122	0.7156	0.7190	0.7223	0.7257	0.7290	0.7323	0.7356
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

DRAFT

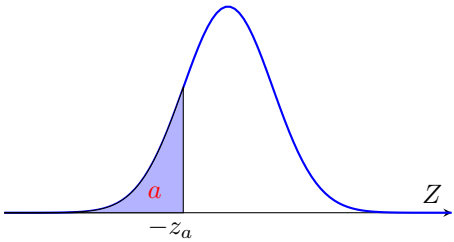


Figure 13.1: Standard Normal Distribution

A.4: The left-sided critical points, $-z_a$, of the standard normal distribution for the probability a.

a	**0**	**0.001**	**0.002**	**0.003**	**0.004**	**0.005**	**0.006**	**0.007**	**0.008**	**0.009**
0.00		- 3.090	-2.878	- 2.748	- 2.652	- 2.576	- 2.512	- 2.457	- 2.409	- 2.366
0.01	-2.326	-2.290	-2.257	-2.226	-2.197	-2.170	-2.144	-2.120	-2.097	-2.075
0.02	-2.054	-2.034	-2.014	-1.995	-1.977	-1.960	-1.943	-1.927	-1.911	-1.896
0.03	-1.881	-1.866	-1.852	-1.838	-1.825	-1.812	-1.799	-1.787	-1.774	-1.762
0.04	-1.751	-1.739	-1.728	-1.717	-1.706	-1.695	-1.685	-1.675	-1.665	-1.655
0.05	-1.645	-1.635	-1.626	-1.616	-1.607	-1.598	-1.589	-1.580	-1.572	-1.563
0.06	-1.555	-1.546	-1.538	-1.530	-1.522	-1.514	-1.506	-1.499	-1.491	-1.483
0.07	-1.476	-1.468	-1.461	-1.454	-1.447	-1.440	-1.433	-1.426	-1.419	-1.412
0.08	-1.405	-1.398	-1.392	-1.385	-1.379	-1.372	-1.366	-1.359	-1.353	-1.347
0.09	-1.341	-1.335	-1.329	-1.323	-1.317	-1.311	-1.305	-1.299	-1.293	-1.287
0.10	-1.282	-1.276	-1.270	-1.265	-1.259	-1.254	-1.248	-1.243	-1.237	-1.232
0.11	-1.227	-1.221	-1.216	-1.211	-1.206	-1.200	-1.195	-1.190	-1.185	-1.180
0.12	-1.175	-1.170	-1.165	-1.160	-1.155	-1.150	-1.146	-1.141	-1.136	-1.131
0.13	-1.126	-1.122	-1.117	-1.112	-1.108	-1.103	-1.098	-1.094	-1.089	-1.085
0.14	-1.080	-1.076	-1.071	-1.067	-1.063	-1.058	-1.054	-1.049	-1.045	-1.041
0.15	-1.036	-1.032	-1.028	-1.024	-1.019	-1.015	-1.011	-1.007	-1.003	-0.999
0.16	-0.994	-0.990	-0.986	-0.982	-0.978	-0.974	-0.970	-0.966	-0.962	-0.958
0.17	-0.954	-0.950	-0.946	-0.942	-0.938	-0.935	-0.931	-0.927	-0.923	-0.919
0.18	-0.915	-0.912	-0.908	-0.904	-0.900	-0.896	-0.893	-0.889	-0.885	-0.882
0.19	-0.878	-0.874	-0.871	-0.867	-0.863	-0.860	-0.856	-0.852	-0.849	-0.845
0.20	-0.842	-0.838	-0.834	-0.831	-0.827	-0.824	-0.820	-0.817	-0.813	-0.810

DRAFT

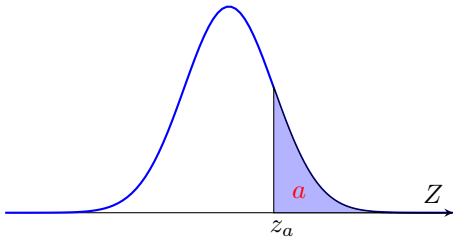


Figure 13.2: Standard Normal Distribution

A.5: The right-sided critical points, z_a, of the standard normal distribution for the probability a

a	**0.000**	**0.001**	**0.002**	**0.003**	**0.004**	**0.005**	**0.006**	**0.007**	**0.008**	**0.009**
0.00		3.090	2.878	2.748	2.652	2.576	2.512	2.457	2.409	2.366
0.01	2.326	2.290	2.257	2.226	2.197	2.170	2.144	2.120	2.097	2.075
0.02	2.054	2.034	2.014	1.995	1.977	1.960	1.943	1.927	1.911	1.896
0.03	1.881	1.866	1.852	1.838	1.825	1.812	1.799	1.787	1.774	1.762
0.04	1.751	1.739	1.728	1.717	1.706	1.695	1.685	1.675	1.665	1.655
0.05	1.645	1.635	1.626	1.616	1.607	1.598	1.589	1.580	1.572	1.563
0.06	1.555	1.546	1.538	1.530	1.522	1.514	1.506	1.499	1.491	1.483
0.07	1.476	1.468	1.461	1.454	1.447	1.440	1.433	1.426	1.419	1.412
0.08	1.405	1.398	1.392	1.385	1.379	1.372	1.366	1.359	1.353	1.347
0.09	1.341	1.335	1.329	1.323	1.317	1.311	1.305	1.299	1.293	1.287
0.10	1.282	1.276	1.270	1.265	1.259	1.254	1.248	1.243	1.237	1.232
0.11	1.227	1.221	1.216	1.211	1.206	1.200	1.195	1.190	1.185	1.180
0.12	1.175	1.170	1.165	1.160	1.155	1.150	1.146	1.141	1.136	1.131
0.13	1.126	1.122	1.117	1.112	1.108	1.103	1.098	1.094	1.089	1.085
0.14	1.080	1.076	1.071	1.067	1.063	1.058	1.054	1.049	1.045	1.041
0.15	1.036	1.032	1.028	1.024	1.019	1.015	1.011	1.007	1.003	0.999
0.16	0.994	0.990	0.986	0.982	0.978	0.974	0.970	0.966	0.962	0.958
0.17	0.954	0.950	0.946	0.942	0.938	0.935	0.931	0.927	0.923	0.919
0.18	0.915	0.912	0.908	0.904	0.900	0.896	0.893	0.889	0.885	0.882
0.19	0.878	0.874	0.871	0.867	0.863	0.860	0.856	0.852	0.849	0.845
0.20	0.842	0.838	0.834	0.831	0.827	0.824	0.820	0.817	0.813	0.810

DRAFT

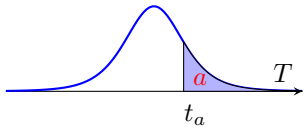
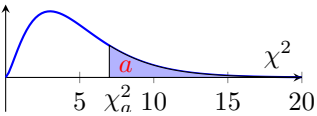


Figure 13.3: Student's-t Distribution

A.6: The right-sided critical points, t_a, of the t-distribution with its degrees of freedom (ν) for the probability a.

ν	a						
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

DRAFT



A.7-: The right-sided critical points, χ^2_a, of the χ^2-distribution with its degrees of freedom (ν) for the probability a.

ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.07	0.12	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

DRAFT

Table 13.2: A.8. Critical Points for $\alpha = 0.05$ of the F-distribution with its degrees of freedom F(df1, df2).

	df1												
df2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58

DRAFT

Table 13.3: Wilcoxon Signed-Rank Test Critical Values for one-tailed test

n	α (one-tailed)					n	α (one-tailed)				
	0.05	0.025	0.01	0.005	0.001		0.05	0.025	0.01	0.005	0.001
5	0	–	–	–	–	20	52	43	37	32	26
6	2	0	–	–	–	21	58	49	42	37	30
7	3	2	0	–	–	22	65	55	48	42	35
8	5	3	1	0	–	23	73	62	54	48	40
9	8	5	3	1	–	24	81	69	61	54	45
10	10	8	5	3	1	25	89	76	68	60	51
11	13	10	7	5	2	26	98	84	75	67	57
12	17	13	9	7	3	27	107	92	83	74	63
13	21	17	12	9	5	28	116	101	91	81	69
14	25	21	15	12	7	29	126	110	100	89	76
15	30	25	19	15	9	30	137	120	109	98	83
16	35	29	23	19	12	31	147	130	118	107	91
17	41	34	27	23	15	32	159	140	128	116	99
18	47	40	32	27	19	33	170	151	138	126	107
19	53	46	37	32	23	34	182	162	148	135	116

Table 13.4: Wilcoxon Signed-Rank Test Critical Values for two-tailed test

n	α (two-tailed)					n	α (two-tailed)				
	0.10	0.05	0.02	0.01	0.002		0.10	0.05	0.02	0.01	0.002
5	0	–	–	–	–	20	60	52	43	37	26
6	2	0	–	–	–	21	67	58	49	42	30
7	3	2	–	–	–	22	75	65	55	48	35
8	5	3	1	–	–	23	83	73	62	54	40
9	8	5	3	1	–	24	91	81	69	61	45
10	10	8	5	3	1	25	100	89	76	68	51
11	13	10	7	5	2	26	110	98	84	75	57
12	17	13	9	7	3	27	119	107	92	83	63
13	21	17	12	9	5	28	130	116	101	91	69
14	25	21	15	12	7	29	140	126	110	100	76
15	30	25	19	15	9	30	151	137	120	109	83
16	35	29	23	19	12	31	163	147	130	118	91
17	41	34	27	23	15	32	175	159	140	128	99
18	47	40	32	27	19	33	187	170	151	138	107
19	53	46	37	32	23	34	200	182	162	148	116

DRAFT

Table 13.5: Mann–Whitney U Critical Values ($\alpha = 0.01$, two-tailed)

n_1	n_2																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	–	–	–	–	–	0	0	0	1	1	1	2	2	2	3	3	4	4	4
4	–	–	–	0	1	1	2	3	3	4	5	5	6	7	8	9	9	10	11
5	–	–	0	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17
6	–	0	1	2	3	5	6	7	9	10	12	13	15	16	18	19	20	22	23
7	–	0	1	3	5	6	8	10	11	13	14	16	18	19	21	23	25	26	28
8	–	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35
9	–	0	3	5	7	10	12	14	17	19	21	24	26	29	31	34	36	39	41
10	–	1	3	6	9	11	14	17	20	23	25	28	31	34	37	40	43	46	49
11	–	1	4	7	10	13	16	19	23	26	29	32	35	38	42	45	48	52	55
12	–	1	5	9	12	14	18	21	25	29	33	36	40	44	47	51	55	59	63
13	–	2	5	10	13	16	20	24	28	32	36	40	44	48	52	56	60	65	69
14	–	2	6	11	15	18	22	26	31	35	40	44	49	53	58	62	67	71	76
15	–	2	7	12	16	19	24	29	34	38	44	48	53	58	63	67	72	77	82
16	–	3	8	13	18	21	26	31	37	42	47	52	58	63	68	73	78	84	89
17	–	3	9	14	19	23	28	34	40	45	51	56	62	67	73	79	84	90	95
18	–	4	9	15	20	25	30	36	43	48	55	60	67	72	78	84	90	96	102
19	–	4	10	16	22	26	32	39	46	52	59	65	71	77	84	90	96	103	109
20	–	4	11	17	23	28	35	41	49	55	63	69	76	82	89	95	102	109	116

DRAFT

Table 13.6: Mann–Whitney U Critical Values ($\alpha = 0.05$, two-tailed)

n_1	n_2																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

DRAFT

Bibliography

[1] David R Anderson, Thomas A Williams, and James J Cochran. *Statistics for Business & Economics*. Cengage Learning, 2020.

[2] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

[3] Peter J Brockwell and Richard A Davis. *Introduction to Time Series and Forecasting*. Springer, 2002.

[4] William S. Cleveland. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1):21–26, 2001.

[5] W. Jay Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, New York, 3rd edition, 1999.

[6] Frederick Emory Croxton and Dudley J Cowden. *Applied General Statistics*. Prentice Hall, 1939.

[7] Wayne W. Daniel. *Applied Nonparametric Statistics*. PWS-Kent, Boston, 2nd edition, 1990.

[8] Anthony J Hayter. *Probability and Statistics for Engineers and Scientists*. Duxbury Press, 2012.

[9] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, New York, 3rd edition, 2014.

[10] Changquan Huang and Alla Petukhina. *Applied Time Series Analysis and Forecasting with Python*. Springer, 2022.

[11] Francesca Lazzeri. *Machine Learning for Time Series Forecasting with Python*. John Wiley & Sons, 2020.

[12] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005. Chapter 5 covers rank tests including Mann-Whitney.

DRAFT

[13] Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. *Forecasting Methods and Applications*. John wiley & sons, 2008.

[14] Robert Deward Mason. *Statistical Techniques in Business and Economics*. McGraw Hill Inc, 1999.

[15] Peter Naur. *Concise Survey of Computer Methods*. Petrocelli/Charter, 1974.

[16] DJ Patil and Jeff Hammerbacher. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(10):70–76, 2012.

[17] Sheldon Ross. *Probability and Statistics for Engineers and Scientists*. Elsevier, New Delhi, 2009.

[18] Sidney Siegel and N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition, 1988.

[19] John W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[20] Ronald E Walpole, Raymond H Myers, Sharon L Myers, and Keying Ye. *Probability and Statistics for Engineers and Scientists*, volume 5. Macmillan New York, 1993.

[21] C.F. Jeff Wu. Statistics = data science?, 1997. Keynote Lecture at the 1997 IMS Annual Meeting, Anaheim, CA.

DRAFT