STATISTICAL INFERENCE



Course code: 8417

Department of Statistics
Faculty of Sciences
Allama Iqbal Open University, Islamabad

STATISTICAL INFERENCE

COURSE CODE: 8417

Department of Statistics
Faculty of Sciences
Allama Iqbal Open University, Islamabad

COURSE DEVELOPMENT TEAM

CHAIRMAN, DEPARTMENT OF STATISTICS:

PROF. DR. IRSHAD AHMAD ARSHAD

COURSE DEVELOPMENT COORDINATOR:

DR. MUHAMMAD YAMEEN DANISH

WRITERS:

DR. MUHAMMAD YAMEEN DANISH

BILAL AHMAD

REVIEWERS:

DR. MUHAMMAD YAMEEN DANISH

BILAL AHMAD

Allama Iqbal Open University, Islamabad.

PREFACE

As the Vice Chancellor of Allama Iqbal Open University, it is my pleasure to introduce this textbook, **Statistical Inference (8417)**. Our institution remains committed to providing accessible, high-quality education that meets the evolving demands of the academic and professional world. Statistics is not merely a technical subject; it is an essential instrument for critical inquiry across all fields of knowledge, enabling us to derive meaning from data and make informed decisions in an increasingly complex society.

The history of Statistics is rooted in the fundamental human desire to understand the patterns of our existence. From the natural sciences to the social sciences, the ability to apply mathematical logic is crucial. As Galileo famously observed, the universe is designed with mathematical precision. This textbook reflects that philosophy by cultivating logical thinking and providing the analytical tools necessary for students to improve living standards and optimize the use of our resources.

The development of this course is a significant step toward empowering our students with essential statistical literacy. The text has been intentionally designed to be clear and accessible, ensuring that complex concepts are approachable without compromising on academic rigor. This performance-based approach, enriched with practical illustrations and deep discussions, is aimed at enhancing both the readability and the conceptual grasp of the subject matter.

I wish to commend the authors and the course development team for their dedicated efforts in producing this comprehensive volume. I am confident that this book will serve as an invaluable resource, providing students with the intellectual satisfaction and visual reinforcement necessary for a profound understanding of statistical inference.

Prof. Dr. Nasir Mahmood
Vice Chancellor
Allama Iqbal Open University Islamabad.

MESSAGE FROM THE CHAIRMAN

Mathematics and Statistics represent the foundational pillars of scientific inquiry, serving as the essential bridge that connects and integrates diverse disciplines within the natural and social sciences. While these fields may initially appear abstract in their theoretical form, their profound utility becomes immediately evident when applied to solve real-world problems and interpret empirical data.

The introduction of Statistical Inference (8417) is a vital addition to our curriculum, specifically designed to instill fundamental quantitative competencies in our students. By mastering these first-level concepts, students develop the analytical rigor necessary to navigate their respective fields with precision. I am pleased to note that this textbook beautifully encapsulates the core principles of both mathematics and statistics, presenting them in a manner that is both comprehensive and pedagogically sound.

I wish to express my sincere appreciation and heartiest congratulations to the course development team and all contributors who have dedicated their expertise to the creation of this work. Their collective effort ensures that our students are equipped with the intellectual tools required for academic excellence and meaningful research.

Prof. Dr. Irshad Ahmad Arshad
Chairman, Department of Statistics
Faculty of Sciences
Allama Iqbal Open University Islamabad.

ACKNOWLEDGEMENTS

I am profoundly grateful to Almighty Allah for the strength and opportunity to lead the development of this course. The successful completion of this textbook is the result of a collective effort, and I wish to express my sincere gratitude to all those who provided their expertise, critical insights, and unwavering support throughout the drafting, editing, and design phases of this project.

I am highly grateful to Prof. Dr. Nasir Mahmood, Vice-Chancellor of AIOU, and Prof. Dr. Irshad Ahmad Arshad, Chairman Department of Statistics, for their visionary leadership and creative guidance. Their commitment to presenting complex mathematical concepts in an accessible manner, along with the provision of all necessary facilities, enabled the completion of this course within a remarkably short period.

I gratefully acknowledge the valuable efforts, support, and thoughtful guidance of Dr. Zahid Majeed, Director, International Collaboration & Exchange, and Director, Academic Planning and Course Production, Allama Iqbal Open University, Islamabad, Pakistan. His keen interest, encouragement, and concern for academic excellence have been instrumental in the successful completion of this book on Statistical Inference (8417).

I am especially indebted to the course writer and composer, Bilal Ahmad, whose dual role was instrumental to this project. His dedication to both the pedagogical content and the technical formatting of the manuscript ensured a high standard of clarity and professionalism.

While every effort has been made to ensure accuracy and clarity, suggestions and constructive criticism for the further improvement of this text are most welcome. Please forward your comments to yameen.danish@aiou.edu.pk.

Dr. Muhammad Yameen Danish
Course Coordinator
Department of Statistics
Allama Iqbal Open University Islamabad.

CONTENTS

UNIT 1: DESCRIPTIVE STATISTICS 1
KEY LEARNING OBJECTIVES 1
1.1 INTRODUCTION 2
A Note on Notation: Greek versus Latin Symbols 2
1.2 MEASURES OF CENTRAL TENDENCY: AN OVERVIEW 3
1.3 THE ARITHMETIC MEAN 3
1.3.1 Arithmetic Mean for Ungrouped Data 4
1.3.2 Arithmetic Mean for Grouped (Frequency-Distribution) Data 4
1.3.3 Important Properties of the Arithmetic Mean 6
1.3.4 Merits and Demerits of the Arithmetic Mean 6
1.4 THE WEIGHTED MEAN 7
1.5 THE GEOMETRIC MEAN 8
1.5.1 Geometric Mean for Ungrouped Data 8
1.5.2 Geometric Mean for Grouped Data 8
1.6 THE MEDIAN 10
1.6.1 Median for Ungrouped Data 10
1.6.2 Median for Grouped Data 11
1.6.3 Merits and Demerits of the Median 12
1.7 THE MODE 12
1.7.1 Mode for Ungrouped Data 12
1.7.2 Mode for Grouped Data 13
1.8 CHOOSING AN APPROPRIATE MEASURE OF CENTRAL TENDENCY 13
1.9 QUARTILES, DECILES AND PERCENTILES 14
1.9.1 Quartiles 14
1.9.2 Deciles and Percentiles 15
1.10 MEASURES OF DISPERSION: AN OVERVIEW 16
1.11 THE RANGE 17
1.11.1 Range for Ungrouped Data 17
1.11.2 Range for Grouped Data 17
1.12 THE QUARTILE DEVIATION (SEMI-INTERQUARTILE RANGE) 18
1.12.1 Quartile Deviation for Ungrouped Data 19
1.12.2 Quartile Deviation for Grouped Data 19
1.13 THE MEAN DEVIATION 20
1.13.1 Mean Deviation for Ungrouped Data 20
1.13.2 Mean Deviation for Grouped Data 20
1.14 THE VARIANCE AND THE STANDARD DEVIATION 22
1.14.1 Variance and Standard Deviation — Definitions 22
1.14.2 Computing Formulas 23

1.14.3 Variance and Standard Deviation for Grouped Data 23
1.15 THE COEFFICIENT OF VARIATION 25
1.16 INTERPRETING THE STANDARD DEVIATION: CHEBYSHEV AND THE EMPIRICAL RULE 26
1.16.1 Chebyshev's Inequality (for Any Distribution) 26
1.16.2 The Empirical Rule (for Bell-Shaped Distributions) 26
1.17 SKEWNESS AND KURTOSIS 27
1.17.1 Skewness 27
1.17.2 Kurtosis 27
1.18 MEASURES OF SKEWNESS AND PEAKEDNESS 28
1.18.1 Karl Pearson's Coefficient of Skewness 28
1.18.2 Bowley's Coefficient of Skewness 28
1.18.3 The Moment Coefficient of Skewness and Kurtosis 29
1.19 SUMMARY 30
1.20 SELF-ASSESSMENT EXERCISE 31
UNIT 2: SAMPLING AND SAMPLING DISTRIBUTIONS 34
KEY LEARNING OBJECTIVES 34
2.1 INTRODUCTION 35
2.2 POPULATION AND SAMPLES 35
2.2.1 Definition of a Population 36
2.2.2 Definition of a Sample 36
2.2.3 The Population–Sample Relationship 37
2.3 PARAMETERS AND ESTIMATES 37
2.3.1 Parameters 37
2.3.2 Statistics and Estimates 38
2.4 INTRODUCTION TO STATISTICAL INFERENCE 39
2.4.1 Estimation 39
2.4.2 Testing of Hypotheses 39
2.5 IMPORTANCE OF SAMPLING IN STATISTICS 40
2.5.1 Reduced Cost 40
2.5.2 Greater Speed 40
2.5.3 Greater Accuracy 40
2.5.4 Greater Scope 40
2.5.5 Necessity 41
2.5.6 When is a Complete Enumeration Preferred? 41
2.6 RANDOM (PROBABILITY) SAMPLING METHODS 41
2.7 SIMPLE RANDOM SAMPLING 41
2.7.1 Sampling With and Without Replacement 42
2.7.2 How to Draw a Simple Random Sample in Practice 43
2.7.3 Advantages and Limitations 43
2.8 STRATIFIED AND PROPORTIONAL STRATIFIED SAMPLING 44

2.8.1 The Idea of Stratification 44
2.8.2 Allocating the Sample to the Strata 44
2.8.3 Advantages of Stratified Sampling 45
2.9 OTHER SAMPLING PROCEDURES 46
2.9.1 Systematic Sampling 46
2.9.2 Cluster Sampling 47
2.9.3 Multi-Stage Sampling 47
2.10 NON-PROBABILITY SAMPLING METHODS 48
2.10.1 Convenience Sampling 48
2.10.2 Quota Sampling 48
2.10.3 Purposive (Judgement) Sampling 49
2.10.4 Snowball Sampling 49
2.11 ERRORS IN SAMPLING 49
2.11.1 Sampling Errors 49
2.11.2 Non-Sampling Errors 49
2.12 THE SAMPLING DISTRIBUTION: A KEY CONCEPT 50
2.12.1 Two Important Things to Remember 50
2.13 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN 50
2.13.1 A Small-Population Illustration 51
2.13.2 General Results for the Sampling Distribution of $\bar{x}$ 52
2.14 THE CENTRAL LIMIT THEOREM 54
2.14.1 Statement of the Central Limit Theorem 54
2.14.2 What "Large n" Means in Practice 54
2.14.3 Why the Central Limit Theorem Matters 54
2.15 THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION 56
2.15.1 Setting Up the Problem 56
2.15.2 Mean and Standard Error of the Sample Proportion 56
2.15.3 Normal Approximation to the Distribution of $\hat{\boldsymbol{p}}$ 57
2.16 THE RELATIONSHIP BETWEEN SAMPLE SIZE AND STANDARD ERROR 58
2.16.1 The Standard Error Decreases as the Sample Size Increases 58
2.16.2 But the Decrease is Slow — the Square Root Effect 58
2.16.3 The Standard Error Does Not Depend on the Population Size 58
2.16.4 Practical Sample-Size Determination 59
2.17 SUMMARY 60
2.18 SELF-ASSESSMENT EXERCISE 61
UNIT 3: ESTIMATION 63
KEY LEARNING OBJECTIVES 63
3.1 INTRODUCTION 64
3.2 POINT ESTIMATION 64
3.3 PROPERTIES OF A GOOD POINT ESTIMATOR 65
3.3.1 Unbiasedness 65

3.3.2 Consistency 66
3.3.3 Efficiency 66
3.3.4 Sufficiency 66
3.4 METHODS OF OBTAINING POINT ESTIMATORS 67
3.4.1 The Method of Moments 67
3.4.2 The Method of Maximum Likelihood 67
3.4.3 The Method of Least Squares 68
3.5 INTERVAL ESTIMATION AND CONFIDENCE INTERVALS 68
3.5.1 The Anatomy of a Confidence Interval 68
3.5.2 Interpreting a Confidence Interval 69
3.6 CONFIDENCE INTERVAL FOR THE POPULATION MEAN (σ KNOWN) 69
3.6.1 Derivation of the Formula 69
3.7 CONFIDENCE INTERVAL FOR THE POPULATION MEAN (σ UNKNOWN): THE t-DISTRIBUTION 71
3.7.1 Properties of the t-Distribution 71
3.7.2 Confidence Interval Formula 71
3.8 CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO MEANS 73
3.8.1 Independent Samples, Population Variances Known 73
3.8.2 Independent Samples, Population Variances Unknown But Assumed Equal 73
3.8.3 Independent Samples, Population Variances Unknown and Unequal 73
3.9 CONFIDENCE INTERVAL FOR A POPULATION PROPORTION 74
3.10 CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS 77
3.11 CONFIDENCE INTERVAL FOR A POPULATION VARIANCE: THE CHI-SQUARE DISTRIBUTION 78
3.11.1 The Chi-Square Distribution 78
3.11.2 Confidence Interval Formula 78
3.12 DETERMINING THE REQUIRED SAMPLE SIZE 79
3.12.1 Sample Size for Estimating a Mean 79
3.12.2 Sample Size for Estimating a Proportion 80
3.13 SUMMARY 81
3.14 SELF-ASSESSMENT EXERCISE 82
UNIT 4: TESTING HYPOTHESES: ONE SAMPLE TESTS 84
KEY LEARNING OBJECTIVES 84
4.1 INTRODUCTION 85
4.2 THE ROLE OF STATISTICAL HYPOTHESIS TESTING 86
4.3 FORMULATING THE NULL AND ALTERNATIVE HYPOTHESES 86
4.3.1 The Null Hypothesis 87
4.3.2 The Alternative Hypothesis 87
4.4 TYPE I AND TYPE II ERRORS 88
4.4.1 Definitions 88
4.4.2 The Trade-Off Between α and β 88

4.5 LEVEL OF SIGNIFICANCE AND CRITICAL REGION 89
4.6 ONE-TAILED AND TWO-TAILED TESTS 89
4.7 GENERAL PROCEDURE FOR A HYPOTHESIS TEST 90
4.8 TESTING THE MEAN OF A NORMAL POPULATION (σ KNOWN): THE Z-TEST 90
4.8.1 The Test Statistic 91
4.9 TESTING THE MEAN OF A NORMAL POPULATION (σ UNKNOWN): THE t-TEST 92
4.9.1 The Test Statistic 93
4.9.2 The Connection Between Confidence Intervals and Hypothesis Tests 95
4.10 TESTING A POPULATION PROPORTION: LARGE SAMPLES 96
4.10.1 The Test Statistic 96
4.11 THE p-VALUE APPROACH 99
4.11.1 Definition of the p-Value 99
4.11.2 The Decision Rule with p-Values 99
4.12 MEASURING THE POWER OF A HYPOTHESIS TEST 100
4.12.1 Computing the Power 101
4.12.2 Power and Sample Size 102
4.13 COMMON PITFALLS AND THEIR REMEDIES 102
4.13.1 Statistical Significance Is Not Practical Importance 102
4.13.2 Failing to Reject Is Not the Same as Accepting 103
4.13.3 Choosing α After Looking at the Data 103
4.13.4 Multiple Testing 103
4.13.5 Confusing the p-Value with the Probability that H_0 Is True 103
4.14 SUMMARY 103
4.15 SELF-ASSESSMENT EXERCISE 104
UNIT 5: TESTING HYPOTHESES: TWO SAMPLE TESTS 106
KEY LEARNING OBJECTIVES 106
5.1 INTRODUCTION 107
5.2 INDEPENDENT VS. DEPENDENT (PAIRED) SAMPLES 108
5.2.1 Independent Samples 108
5.2.2 Dependent (Paired) Samples 108
5.2.3 Why the Distinction Matters 109
5.3 THE SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO MEANS 109
5.4 TESTING THE DIFFERENCE BETWEEN TWO MEANS: LARGE SAMPLES (Z-TEST) 110
5.4.1 The Hypotheses 110
5.4.2 The Test Statistic 110
5.5 TESTING THE DIFFERENCE BETWEEN TWO MEANS: SMALL SAMPLES, EQUAL VARIANCES (POOLED t-TEST) 111
5.5.1 The Assumptions 112
5.5.2 The Pooled Variance and the Test Statistic 112
5.6 TESTING THE DIFFERENCE BETWEEN TWO MEANS: SMALL SAMPLES, UNEQUAL VARIANCES (WELCH'S t-TEST) 113

5.6.1 The Test Statistic 114
5.7 TESTING THE DIFFERENCE BETWEEN TWO MEANS WITH DEPENDENT (PAIRED) SAMPLES 115
5.7.1 The Test Statistic 115
5.8 TESTING THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS 117
5.8.1 The Sampling Distribution 117
5.8.2 The Pooled and Unpooled Standard Errors 117
5.9 THE p-VALUE REVISITED AND THE TWO-SAMPLE DECISION RULE 120
5.10 CHOOSING THE RIGHT TWO-SAMPLE TEST: A DECISION MAP 121
5.11 COMMON PITFALLS IN TWO-SAMPLE TESTING 121
5.11.1 Treating Paired Data as Independent (or Vice Versa) 122
5.11.2 Ignoring Unequal Variances 122
5.11.3 Comparing Samples from Different Populations 122
5.11.4 Multiple Comparisons in Two-Sample Studies 122
5.12 SUMMARY 122
5.13 SELF-ASSESSMENT EXERCISE 123
UNIT 6: CHI-SQUARE AND ANALYSIS OF VARIANCE 127
KEY LEARNING OBJECTIVES 127
6.1 INTRODUCTION 128
6.2 THE CHI-SQUARE DISTRIBUTION: A BRIEF REVIEW 128
6.3 CHI-SQUARE AS A TEST OF GOODNESS OF FIT 129
6.3.1 The Test Statistic 129
6.3.2 The Decision Rule 129
6.4 CHI-SQUARE AS A TEST OF INDEPENDENCE 131
6.4.1 Setting Up the Problem 132
6.4.2 The Test Statistic and Degrees of Freedom 132
6.5 CHI-SQUARE AS A TEST OF HOMOGENEITY 133
6.6 INFERENCE ABOUT A SINGLE POPULATION VARIANCE 135
6.6.1 The Test Statistic 135
6.7 INFERENCE ABOUT TWO POPULATION VARIANCES: THE F-DISTRIBUTION 136
6.7.1 The F-Distribution 136
6.7.2 The Test Statistic 137
6.8 ANALYSIS OF VARIANCE: AN INTRODUCTION 138
6.9 THE ONE-WAY ANOVA MODEL AND NOTATION 138
6.9.1 The Model Assumptions 139
6.9.2 The Hypotheses 139
6.10 COMPUTING THE ONE-WAY ANOVA: SST, SSB, SSW AND THE F-RATIO 139
6.10.1 Total Sum of Squares (SST) 140
6.10.2 Sum of Squares Between Groups (SSB) 140
6.10.3 Sum of Squares Within Groups (SSW) 140
6.10.4 The Fundamental Identity 140

6.10.5 Mean Squares and the F-Ratio 140
6.11 THE ANOVA TABLE AND THE DECISION RULE 141
6.12 ASSUMPTIONS OF ANOVA AND WHAT TO DO WHEN THEY FAIL 143
6.12.1 Violations of Independence 143
6.12.2 Violations of Normality 143
6.12.3 Violations of Homogeneity of Variances 143
6.13 COMMON PITFALLS IN CHI-SQUARE AND ANOVA 143
6.13.1 Small Expected Frequencies in Chi-Square 144
6.13.2 Confusing Independence with Homogeneity 144
6.13.3 Treating ANOVA Rejection as Identification 144
6.13.4 Ignoring Heterogeneity of Variances in ANOVA 144
6.14 SUMMARY 144
6.15 SELF-ASSESSMENT EXERCISE 145
UNIT 7: REGRESSION AND CORRELATION ANALYSIS 149
KEY LEARNING OBJECTIVES 149
7.1 INTRODUCTION 150
7.2 SCATTER DIAGRAMS AND THE IDEA OF A LINEAR RELATIONSHIP 150
7.3 THE SIMPLE LINEAR REGRESSION MODEL 151
7.3.1 The Assumptions (LINE) 151
7.4 ESTIMATION OF THE REGRESSION COEFFICIENTS BY THE METHOD OF LEAST SQUARES 152
7.5 INTERPRETING THE REGRESSION EQUATION: SLOPE AND INTERCEPT 153
7.6 THE STANDARD ERROR OF ESTIMATE 154
7.7 THE COEFFICIENT OF DETERMINATION (R^2) 154
7.8 CORRELATION ANALYSIS: THE PEARSON CORRELATION COEFFICIENT 155
7.8.1 The Sample Correlation Coefficient 155
7.8.2 Properties of r 155
7.8.3 Interpreting the Magnitude of r 157
7.9 INFERENCES ABOUT POPULATION REGRESSION PARAMETERS 158
7.9.1 The ANOVA Approach to Testing the Overall Regression 158
7.9.2 The Standard Error of the Slope 158
7.9.3 Test of Hypothesis for the Slope 159
7.9.4 Confidence Interval for the Slope 159
7.9.5 Test and Confidence Interval for the Correlation Coefficient 159
7.10 PREDICTION USING THE REGRESSION LINE 160
7.11 LIMITATIONS, ERRORS AND CAVEATS OF REGRESSION AND CORRELATION 161
7.11.1 Correlation Does Not Imply Causation 161
7.11.2 Do Not Extrapolate Beyond the Range of the Data 161
7.11.3 Beware of Influential Observations and Outliers 161
7.11.4 $r \approx 0$ Does Not Mean No Relationship 162
7.11.5 Ecological Fallacy 162

7.11.6 Regression Towards the Mean 162
7.11.7 The Spearman Rank Correlation Coefficient 162
7.12 MULTIPLE REGRESSION AND CORRELATION ANALYSIS 163
7.13 FINDING THE MULTIPLE REGRESSION EQUATION 164
7.14 INFERENCES IN MULTIPLE REGRESSION 165
7.14.1 The Overall F-Test 165
7.14.2 Individual t-Tests for Each Coefficient 165
7.14.3 The Adjusted R^2 165
7.15 MODELLING TECHNIQUES: VARIABLE SELECTION AND DIAGNOSTICS 165
7.15.1 Variable Selection 166
7.15.2 Multicollinearity 166
7.15.3 Residual Diagnostics 166
7.16 SUMMARY 166
7.17 SELF-ASSESSMENT EXERCISE 167
UNIT 8: NONPARAMETRIC METHODS 171
KEY LEARNING OBJECTIVES 171
8.1 INTRODUCTION 172
8.2 WHEN ARE NONPARAMETRIC METHODS NEEDED? 172
8.2.1 Severe violation of normality 172
8.2.2 Presence of outliers 173
8.2.3 Ordinal (ranked) data 173
8.2.4 Very small sample sizes 173
8.2.5 When the parameter of interest is the median rather than the mean 173
8.2.6 When the population variance is not constant (heteroscedasticity) 173
8.2.7 Relationship between nonparametric and parametric tests 173
8.3 THE SIGN TEST FOR A POPULATION MEDIAN 174
8.3.1 Derivation from the Binomial Distribution 174
8.3.2 Handling Ties and Zero Differences 175
8.3.3 Normal Approximation for Large Samples 175
8.3.4 One-Sided and Two-Sided Tests 175
8.4 THE SIGN TEST FOR PAIRED DATA 176
8.5 THE WILCOXON SIGNED-RANK TEST FOR PAIRED DATA 177
8.5.1 Theoretical Foundation: Why Ranks? 177
8.5.2 Exact Distribution of W^+ 177
8.5.3 Mean and Variance under H_0 178
8.5.4 Normal Approximation and Continuity Correction 178
8.5.5 Ties and Their Effect 178
8.6 THE MANN-WHITNEY U TEST (WILCOXON RANK-SUM TEST) 179
8.6.1 Conceptual Basis: Stochastic Superiority 179
8.6.2 The U Statistic and Its Relationship to W 180
8.6.3 Mean and Variance under H_0 180

8.6.4 Normal Approximation 180
8.6.5 Handling Ties in the Rank-Sum Test 180
8.7 THE KRUSKAL-WALLIS H TEST FOR SEVERAL INDEPENDENT SAMPLES 182
8.7.1 Extension of the Mann-Whitney to k > 2 182
8.7.2 The H Statistic and Its Chi-Square Approximation 182
8.7.3 Post-Hoc Comparisons after a Significant H 182
8.7.4 Assumptions and Robustness 182
8.8 THE ONE-SAMPLE RUNS TEST FOR RANDOMNESS 184
8.8.1 The Concept of a Run 184
8.8.2 Exact Distribution of the Number of Runs 185
8.8.3 Mean and Variance under Randomness 185
8.8.4 Normal Approximation 185
8.8.5 Using the Runs Test with Continuous Data (Above/Below Median) 185
8.9 THE SPEARMAN RANK CORRELATION TEST 186
8.9.1 From Pearson to Spearman: The Rank Transformation 186
8.9.2 Derivation of the Simplified Formula 186
8.9.3 Ties and the Exact Calculation 186
8.9.4 Hypothesis Test for ρ_s 187
8.9.5 Relationship to Kendall's Tau 187
8.10 THE KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST 187
8.10.1 The Empirical Distribution Function (ECDF) 188
8.10.2 The One-Sample K-S Statistic 188
8.10.3 Critical Values and the Lilliefors Correction 188
8.10.4 The Two-Sample K-S Test 188
8.11 FACTORS AFFECTING THE MAGNITUDE OF CORRELATION 188
8.11.1 Range restriction 188
8.11.2 Outliers 188
8.11.3 Non-linearity 189
8.11.4 Heteroscedasticity 189
8.11.5 Measurement error 189
8.11.6 Sample size 189
8.12 COMPARING PARAMETRIC AND NONPARAMETRIC TESTS: STRENGTHS AND WEAKNESSES 189
8.12.1 Asymptotic Relative Efficiency 189
8.12.2 Power Comparisons under Different Distributions 189
8.12.3 Advantages of Nonparametric Tests 189
8.12.4 Disadvantages 190
8.13 CHOOSING THE RIGHT NONPARAMETRIC TEST: A DECISION MAP 190
8.13.1 How many samples? 190
8.13.2 Type of question? 190
8.13.3 Check assumptions: 190

8.14 COMMON PITFALLS IN NONPARAMETRIC TESTING 190
8.14.1 Using a nonparametric test when a parametric test is clearly appropriate 190
8.14.2 Ignoring ties 190
8.14.3 Misinterpreting the null hypothesis 191
8.14.4 Small sample sizes without exact p-values 191
8.14.5 Using the runs test on non-dichotomous data without a valid splitting rule 191
8.14.6 Over-reliance on p-values without effect size 191
8.14.7 Applying nonparametric tests to data with very small sample sizes (n < 5) without caution 191
8.15 SUMMARY 191
8.16 SELF-ASSESSMENT EXERCISE 192
UNIT 9: TIME SERIES & FORECASTING 195
KEY LEARNING OBJECTIVES 195
9.1 INTRODUCTION 196
9.2 THE MEANING AND USES OF TIME SERIES ANALYSIS 196
9.3 THE FOUR COMPONENTS OF A TIME SERIES 197
9.3.1 Secular Trend (T) 197
9.3.2 Cyclical Variation (C) 197
9.3.3 Seasonal Variation (S) 197
9.3.4 Irregular Variation (I) 197
9.4 TIME SERIES DECOMPOSITION MODELS: ADDITIVE AND MULTIPLICATIVE 197
9.4.1 The Additive Model 197
9.4.2 The Multiplicative Model 198
9.4.3 Which Model to Choose? 198
9.5 MEASURING THE SECULAR TREND: AN OVERVIEW 198
9.6 THE FREEHAND (GRAPHICAL) METHOD OF TREND MEASUREMENT 198
9.6.1 The Procedure 198
9.6.2 Limitations of the Freehand Method 199
9.7 THE METHOD OF SEMI-AVERAGES FOR TREND MEASUREMENT 200
9.7.1 The Procedure 200
9.7.2 Merits and Demerits 200
9.8 THE METHOD OF MOVING AVERAGES FOR TREND MEASUREMENT 201
9.8.1 The Procedure 201
9.8.2 Choosing the Order of the Moving Average 201
9.9 THE METHOD OF LEAST SQUARES FOR TREND MEASUREMENT 203
9.9.1 Fitting a Linear Trend 203
9.9.2 The Standard Error of Estimate 203
9.10 MEASURING SEASONAL VARIATIONS: AN OVERVIEW 205
9.11 THE RATIO-TO-TREND METHOD FOR MEASURING SEASONAL VARIATIONS 205
9.11.1 Procedure 205
9.11.2 Interpreting Seasonal Indices 206

9.12 THE RATIO-TO-MOVING-AVERAGE METHOD FOR MEASURING SEASONAL VARIATIONS 207
9.12.1 Procedure 207
9.12.2 Advantages of the Ratio-to-Moving-Average Method 207
9.13 DESEASONALIZING A TIME SERIES 208
9.13.1 Procedure for Deseasonalization 208
9.13.2 Why Deseasonalize? 208
9.14 MEASURING CYCLICAL VARIATION 208
9.14.1 The Residual Method 209
9.14.2 Interpreting the Cyclical Component 209
9.15 FORECASTING WITH TIME SERIES MODELS 209
9.15.1 Forecasting with the Multiplicative Model 209
9.15.2 Forecasting with the Additive Model 209
9.15.3 Limitations of Time Series Forecasting 209
9.16 EXPONENTIAL SMOOTHING METHODS FOR FORECASTING 210
9.16.1 Simple Exponential Smoothing (SES) 210
9.16.2 Holt's Linear Exponential Smoothing (Double Exponential Smoothing) 210
9.16.3 Holt-Winters Exponential Smoothing (Triple Exponential Smoothing) 210
9.17 CHOOSING THE RIGHT FORECASTING METHOD: A DECISION MAP 211
9.18 COMMON PITFALLS IN TIME SERIES ANALYSIS 211
9.18.1 Extrapolating Too Far Beyond the Data 211
9.18.2 Ignoring the Seasonal Pattern 212
9.18.3 Using an Inappropriate Decomposition Model 212
9.18.4 Confusing the Cyclical Component with the Trend 212
9.18.5 Overfitting the Model 212
9.18.6 Forgetting that Correlation is Not Causation 212
9.19 SUMMARY 212
9.20 SELF-ASSESSMENT EXERCISE 213
REFERENCES AND SUGGESTED READINGS 216
APPENDIX: STATISTICAL TABLES 217
TABLE 1: STANDARD NORMAL DISTRIBUTION 218
TABLE 2: STUDENT'S t-DISTRIBUTION 219
TABLE 3: CHI-SQUARE (χ^2) DISTRIBUTION 220
TABLE 4: WILCOXON SIGNED-RANK TEST 221
TABLE 5: SPEARMAN RANK CORRELATION COEFFICIENT 222
TABLE 6: KOLMOGOROV-SMIRNOV TEST 223
TABLE 7: PEARSON CORRELATION COEFFICIENT 224
TABLE 8: F-DISTRIBUTION CRITICAL VALUES ($\boldsymbol{\alpha = 0.05}$) 225
TABLE 9: F-DISTRIBUTION CRITICAL VALUES ($\alpha = 0.01$) 226
TABLE 10: MANN-WHITNEY U TEST CRITICAL VALUES 227

UNIT 1: DESCRIPTIVE STATISTICS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Distinguish clearly between a population and a sample, and use the correct Greek (population) and Latin (sample) notation when reporting descriptive measures.
2. Compute and interpret the principal measures of central tendency — the arithmetic, weighted and geometric means, the median and the mode — for both ungrouped and grouped data, and decide which of them is most appropriate in a given practical situation.
3. Calculate quartiles, deciles and percentiles for both ungrouped and grouped data, and use them to describe the relative position of individual observations within a distribution.
4. Compute and interpret the most widely used measures of dispersion (range, quartile deviation, mean deviation, variance, standard deviation) and the coefficient of variation, and select the appropriate measure for a particular type of data.
5. Recognize the shape of a distribution by computing measures of skewness and kurtosis, and use these together with the measures of central tendency and dispersion to write a complete numerical description of a data set drawn from any social-science context.

1.1 INTRODUCTION

Statistics is the science of collecting, organizing, presenting, analyzing and interpreting numerical information so that sound decisions can be made in the face of uncertainty. In the social sciences — economics, sociology, psychology, education, public administration and business — researchers are seldom able to examine every member of the group they are interested in. The total cost, the time involved, and the practical impossibility of reaching everyone usually make a complete enumeration impossible. Instead, the researcher works with a small but carefully chosen part of the group, called a sample, and uses the information contained in the sample to draw conclusions about the larger group, called the population. The branch of statistics that draws such conclusions is called inferential statistics, and it is the central concern of this entire course.

Before any inference can be drawn, however, the raw data must first be summarized in a meaningful way. A list of three hundred examination marks, five hundred household incomes, or two thousand replies to an opinion survey is, in its raw form, almost impossible for the human mind to grasp. We need compact numerical quantities that capture the most important features of the data. These quantities are the subject of the present unit, and are collectively known as descriptive statistics. They form the foundation on which all the more advanced inferential techniques in later units of this book are built. Without a firm grasp of descriptive statistics, the ideas of estimation, hypothesis testing, regression and forecasting that follow will remain obscure.

Descriptive statistics answer four very natural questions about a data set:

1. Where is the centre of the data? This question is answered by the measures of central tendency — the mean, the median and the mode.
2. How widely are the observations spread around that centre? This question is answered by the measures of dispersion — the range, the quartile deviation, the mean deviation, the variance, the standard deviation and the coefficient of variation.
3. Where does a particular observation stand in relation to the rest of the data? This question is answered by the quantiles, namely the percentiles, deciles and quartiles.
4. What is the overall shape of the distribution — is it symmetric or lop-sided, sharply peaked or rather flat? This question is answered by the measures of skewness and kurtosis.

The remainder of this unit takes up these four questions in turn. Each new measure is first defined in plain words, then expressed mathematically by means of a formula (often more than one — a definitional form, a computing form, and sometimes a short-cut form), and finally illustrated by at least one solved numerical example for ungrouped data and another for grouped data. We strongly advise the student to work through every example with paper and pencil before attempting the exercises at the end of the unit. Statistics, like swimming, cannot be learned by reading alone.

A Note on Notation: Greek versus Latin Symbols

Throughout this unit, and indeed throughout the entire course, we follow the universal statistical convention of reserving Greek letters for population parameters and Latin letters for sample statistics. This convention is so important that students should commit it to memory at the very outset. The most frequently encountered symbols are summarized in the table below.

Quantity	Population Parameter (Greek)	Sample Statistic (Latin)
Number of observations	N	n

Quantity	Population Parameter (Greek)	Sample Statistic (Latin)
Mean	μ (mu)	$\bar{x}$ (x-bar)
Variance	σ^2 (sigma-squared)	s^2
Standard deviation	σ (sigma)	s
Proportion	π (pi) or P	$\hat{p}$ or p
Correlation coefficient	ρ (rho)	r

In real-world research we almost always work with a sample, and so we compute the Latin-letter quantities directly from the data. These sample statistics are then used to estimate the unknown Greek-letter parameters of the population from which the sample was drawn. Because the present course is devoted to statistical inference, students will encounter both kinds of formulas repeatedly. In this unit we have made a deliberate effort, wherever possible, to write the sample and population formulas side by side, so that students develop an instinct for the distinction from the very first chapter. This habit, once acquired, will pay dividends in every subsequent unit.

1.2 MEASURES OF CENTRAL TENDENCY: AN OVERVIEW

A measure of central tendency is a single number that describes the centre or the typical value of a data set. It is sometimes loosely called an average or a representative value. The idea behind a central-tendency measure is intuitive: when we are presented with hundreds of observations, we naturally look for one number that captures the "essence" of the whole data set, around which the individual observations seem to cluster. A salary survey of one thousand employees, for instance, can be summarized by saying that the average monthly salary is, say, Rs. 65,000. That single figure conveys a lot of information very quickly.

A good measure of central tendency should ideally satisfy several properties: it should be rigidly defined so that two researchers analyzing the same data arrive at the same answer; it should be easy to compute and to understand; it should be based on every observation in the data set; it should be reasonably stable from sample to sample, so that small changes in the data do not produce wild changes in the value of the measure; and it should lend itself to further mathematical and statistical treatment, so that it can serve as a building block in more advanced techniques. As we shall see, no single measure satisfies all these properties perfectly, and that is why we need more than one.

The three most widely used measures of central tendency are the arithmetic mean, the median and the mode. In addition, two specialized forms of the mean — the weighted mean and the geometric mean — are used in particular situations. Each of these answers the question "where is the centre of the data?" in a slightly different way, and the choice between them depends on both the nature of the data and the purpose of the investigation. We shall study each in turn in Sections 1.3 to 1.7, and then discuss in Section 1.8 how to decide which of them to use in any given practical situation.

1.3 THE ARITHMETIC MEAN

The arithmetic mean is the most familiar measure of central tendency. It is what most people have in mind when they use the everyday word "average". The arithmetic mean is obtained by adding together all the observations in the data set and dividing the total by the number of observations. Despite its simplicity, the arithmetic mean is enormously important in statistical theory because it lends itself readily to algebraic manipulation, and because nearly every advanced inferential technique — confidence intervals, tests of hypotheses, regression analysis, analysis of variance — is built on it in one way or another.

1.3.1 Arithmetic Mean for Ungrouped Data

Suppose we have a sample of n observations denoted by $x_1, x_2, \ldots, x_n$ drawn from a population that consists of N observations $X_1, X_2, \ldots, X_N$. The sample mean is denoted by $\bar{x}$ (read as "x-bar") and the population mean is denoted by the Greek letter μ. They are defined side by side as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (1.1)$$

In these formulas, the symbol Σ (the capital Greek letter sigma) means "the sum of". The expression Σx_i from i = 1 to n therefore stands for the sum $x_1 + x_2 + \cdots + x_n$. The symbols x_i and X_i are the i-th observations in the sample and the population respectively, while n and N are the sample size and the population size. Notice that the two formulas are mathematically identical in form: only the symbols differ. The left-hand formula uses the lower-case Latin letter x and the divisor n (the sample size), while the right-hand formula uses the upper-case Latin letter X with the Greek letter μ and the divisor N (the population size). The student should pause and re-read this paragraph until the distinction is completely clear, because the same Latin-versus-Greek distinction will appear again and again throughout the unit.

Example 1.1 (Arithmetic Mean — Ungrouped Sample Data)

The marks obtained by ten students of BS Sociology at Allama Iqbal Open University in the mid-term examination of "Research Methods" are recorded as 68, 72, 55, 81, 66, 74, 59, 85, 70 and 60. Compute the arithmetic mean of these marks.

Solution: Because these ten students form a sample drawn from the much larger population of all BS Sociology students at the university, we use the sample formula. We first compute the sum of the observations:

$$\sum_{i=1}^{10} x_i = 68 + 72 + 55 + 81 + 66 + 74 + 59 + 85 + 70 + 60 = 690.$$

The sample size is n = 10. Therefore, applying formula (1.1):

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{690}{10} = 69\ marks.$$

The typical student in this sample scored 69 marks out of 100 in the examination. Notice how a single number — 69 — has summarized the performance of all ten students in a very compact and informative way.

1.3.2 Arithmetic Mean for Grouped (Frequency-Distribution) Data

When the data set is large, it is usually presented in the form of a frequency distribution rather than as a long list of individual values. In a frequency distribution, the observations are grouped into a number of classes (also called class intervals), and instead of listing every observation, only the class limits and the corresponding frequencies (the number of observations falling in each class) are reported. This is a much more compact way of summarizing large data sets, but it has one drawback: once we have grouped the data, we no longer know the exact value of any individual observation. We only know that, say, 18 observations fell somewhere in the class 50 – 59. To compute statistics from such grouped data, we make a simplifying assumption: every observation in a class is taken to be equal to the mid-point of that class. The mid-point of

the class 50 – 59 (with class boundaries 49.5 – 59.5) is 54.5, and so we treat all 18 observations in this class as if they were equal to 54.5.

Let $x_1, x_2, \ldots, x_k$ denote the mid-points of the k classes and let $f_1, f_2, \ldots, f_k$ denote the corresponding class frequencies, with $\Sigma f_i = n$ (or N, in the population case). Then the sample and population means for grouped data are:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \qquad \mu = \frac{\sum_{i=1}^{k} f_i X_i}{N} \qquad (1.2)$$

When the class mid-points are large or awkward, the arithmetic involved in directly multiplying out $f_i x_i$ for every class can be heavy. A short-cut, called the coding method or the change-of-origin-and-scale method, makes the computation considerably easier. We choose a provisional mean a (sometimes called an assumed mean), preferably the mid-point of one of the central classes, and we let h be the common class width. Define the coded variable:

$$u_i = \frac{x_i - a}{h} \qquad (1.3)$$

The values of u_i are usually small whole numbers $(\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots)$, so the multiplications $f_i u_i$ are very easy to perform. The mean is then recovered from the coded values by:

$$\bar{x} = a + h \cdot \frac{\sum_{i=1}^{k} f_i u_i}{\sum_{i=1}^{k} f_i} \qquad (1.4)$$

Both methods — the direct method based on equation (1.2) and the coding method based on equation (1.4) — produce exactly the same value of $\bar{x}$, except for very small differences caused by rounding. A student who is comfortable with both methods is well prepared for the kinds of computational questions that are typical in examinations and in real research.

Example 1.2 (Arithmetic Mean — Grouped Data, Direct and Coding Methods)

The following frequency distribution shows the monthly household expenditure (in thousand rupees) of 60 families chosen at random from a residential neighborhood in Rawalpindi:

Expenditure	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
No. of Families	4	9	14	18	8	5	2

Compute the average monthly household expenditure (a) by the direct method and (b) by the coding method.

Solution: We begin by preparing a calculation table. The class mid-points are obtained by averaging the lower and upper class limits of each class. We then construct the columns fx, u and fu, taking the provisional mean a = 54.5 (the mid-point of the central class) and the class width h = 10.

Class	f	X	fx	u = (x − 54.5)/10	fu
20 – 29	4	24.5	98.0	−3	−12
30 – 39	9	34.5	310.5	−2	−18
40 – 49	14	44.5	623.0	−1	−14
50 – 59	18	54.5	981.0	0	0
60 – 69	8	64.5	516.0	1	8

70 – 79	5	74.5	372.5	2	10
80 – 89	2	84.5	169.0	3	6
Total	60	—	3070.0	—	−20

(a) Direct method. Applying equation (1.2):

$$\bar{x} = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{3070.0}{60} = 51.17 \; thousand \; rupees.$$

(b) Coding method. Applying equation (1.4) with a = 54.5 and h = 10:

$$\bar{x} = 54.5 + 10 \cdot \frac{\Sigma fu}{\Sigma f} = 54.5 + 10 \cdot \frac{(-20)}{60} = 54.5 - 3.33 = 51.17 \; thousand \; rupees.$$

Both methods agree perfectly: the average monthly household expenditure in this neighborhood of Rawalpindi is approximately Rs. 51,170. The coding method involved much smaller numbers (the largest absolute value in the fu column is 18, compared with 981.0 in the fx column), and so it is the safer choice when calculations are being done by hand.

1.3.3 Important Properties of the Arithmetic Mean

The arithmetic mean possesses a number of algebraic properties that make it especially useful in statistical theory. Five of these are worth memorizing.

1. The sum of the deviations of the observations from their mean is always exactly zero. In symbols, $\Sigma(x_i - \bar{x}) = 0$. This property captures the idea that the mean is a centre of balance: the positive deviations on the right of the mean are exactly cancelled by the negative deviations on the left.
2. The sum of the squared deviations from the mean is a minimum. In other words, $\Sigma(x_i - \bar{x})^2$ is smaller than $\Sigma(x_i - c)^2$ for any other constant c. This least-squares property is the deepest mathematical reason why the mean occupies such a central place in inferential statistics — it is the value that "best fits" the data in the sense of minimum total squared error.
3. If every observation in the data set is multiplied by a constant k, the mean is also multiplied by k. If every observation is increased (or decreased) by a constant k, the mean is also increased (or decreased) by k. These two facts together describe how the mean responds to a change of scale and a change of origin.
4. When two samples of sizes n_1 and n_2 with means $\bar{x}_1$ and $\bar{x}_2$ are combined into a single larger sample of size $n_1 + n_2$, the mean of the combined sample is given by the weighted formula $\bar{x}_c = (n_1\bar{x}_1 + n_2\bar{x}_2) / (n_1 + n_2)$. This is sometimes called the combined mean or pooled mean.
5. The arithmetic mean is sensitive to every observation in the data set. Adding or removing even a single observation will, in general, change the value of the mean.

1.3.4 Merits and Demerits of the Arithmetic Mean

Merits. The arithmetic mean is rigidly defined; it is easy to compute and easy to understand; it is based on every observation in the data set; it is amenable to further mathematical treatment; and it is the most stable

of all the measures of central tendency from sample to sample. For these reasons it is the foundation of most inferential techniques.

Demerits. The mean is highly sensitive to extreme observations (outliers). A single unusually large or small value can drag the mean far away from the bulk of the data. For example, the average wealth of ten ordinary people in a room becomes wildly misleading if a billionaire walks in. The mean cannot be computed for open-ended class intervals (such as "80 and above") without making additional assumptions, and it cannot be used for purely qualitative data such as eye colour or marital status. Finally, the mean may itself take a value that does not actually occur in the data set: a family may have a mean of 2.4 children, but no family actually has 2.4 children.

1.4 THE WEIGHTED MEAN

The ordinary arithmetic mean treats every observation as equally important. There are, however, many practical situations in which some observations deserve more importance than others. For example, when computing a student's grade-point average (GPA), a course worth four credit hours should count more than a course worth only two. When computing the average price of a basket of consumer goods, those that are consumed in larger quantities should count more than those consumed in smaller quantities. When computing the average rate of return on a portfolio of investments, the investments with larger amounts of money tied up in them should count more than the smaller ones. In all such situations the appropriate measure of centre is the weighted arithmetic mean.

Let $x_1, x_2, \ldots, x_n$ be the observations and let $w_1, w_2, \ldots, w_n$ be the corresponding weights. The weights reflect the relative importance assigned to each observation. The weighted mean, denoted by $\bar{x}_w$, is then defined as:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (1.5)$$

When all the weights are equal — say, when every w_i equals 1 — the weighted mean reduces to the ordinary arithmetic mean. The weighted mean is therefore a generalization of the simple arithmetic mean, not a different concept altogether. In choosing the weights, the researcher must use judgement guided by the nature of the problem. Credit hours, quantities consumed, market shares, populations of cities, and amounts of money invested are all common choices in different applications.

Example 1.3 (Weighted Mean — Grade Point Average)

A student of BS Economics at AIOU took five courses in a particular semester. The credit hours and the grade points awarded in each course are shown below. Compute the student's semester GPA.

Course	Credit Hours (w)	Grade Points (x)	wx
Microeconomics	3	3.7	11.10
Statistical Inference	3	3.3	9.90
Research Methods	3	4.0	12.00
English	2	2.7	5.40
Islamic Studies	2	3.0	6.00
Total	13	—	44.40

Solution. Applying the weighted-mean formula (1.5):

$$GPA = \frac{\Sigma wx}{\Sigma w} = \frac{44.40}{13} = 3.42.$$

The semester GPA of this student is 3.42. For comparison, the simple arithmetic mean of the five grade points would have been (3.7 + 3.3 + 4.0 + 2.7 + 3.0)/5 = 3.34, which understates the student's actual academic performance because it ignores the fact that the higher grades were earned in courses with the higher credit-hour weights. This example shows very clearly that ignoring the weights, when weights are present, can give a misleadingly low (or high) answer.

1.5 THE GEOMETRIC MEAN

The geometric mean is the n-th root of the product of n positive observations. It is particularly useful when the data represent rates of change, ratios, percentages or index numbers — situations in which the underlying quantities are multiplied together rather than added. Examples include the average annual growth rate of a population, the average rate of inflation over several years, the average compound rate of return on an investment, and the average rate of growth of a business. In all these situations the arithmetic mean gives a slightly wrong answer; the geometric mean gives the right one. It should be noted at the outset that the geometric mean is defined only for positive observations, because the product of negative or zero values cannot meaningfully be raised to a fractional root.

1.5.1 Geometric Mean for Ungrouped Data

For a sample of n positive observations $x_1, x_2, \ldots, x_n$ the geometric mean (GM) is defined by:

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} \quad (nth\ root) \qquad (1.6)$$

In practice, computing an n-th root by hand is awkward for any sizeable n, so it is usually more convenient to compute the geometric mean indirectly by taking logarithms. Taking logarithms of both sides of equation (1.6) gives the practical formula:

$$\log(GM) = \frac{\sum_{i=1}^{n} \log x_i}{n} \quad , \quad so\ that\ GM = antilog\left(\frac{\Sigma \log x}{n}\right) \qquad (1.7)$$

In words: the logarithm of the geometric mean is the arithmetic mean of the logarithms. This pleasant property makes the geometric mean very easy to compute as soon as a table of logarithms or a scientific calculator is available.

1.5.2 Geometric Mean for Grouped Data

When the data are presented in the form of a frequency distribution with mid-points $x_1, x_2, \ldots, x_k$ and frequencies $f_1, f_2, \ldots, f_k$, the geometric mean is computed by:

$$\log(GM) = \frac{\sum_{i=1}^{k} f_i \log x_i}{\sum_{i=1}^{k} f_i} \qquad (1.8)$$

and then GM is recovered by taking the antilog. The reasoning is identical to the ungrouped case, except that each $\log x_i$ is now multiplied by the frequency f_i before being summed.

Example 1.4 (Geometric Mean — Ungrouped Data, Growth Rates)

The annual rates of growth of a small-scale industry in Pakistan during five consecutive years were 5%, 8%, 12%, 7% and 10%. Find the average annual growth rate.

Solution. Because growth rates compound (multiply) rather than simply add, the appropriate measure of centre is the geometric mean. We must, however, work not with the percentages themselves but with the corresponding "growth factors" 1 + r, since a growth rate of 5% means that the quantity is multiplied by 1.05.

Year	Rate	Factor x = 1 + r	log x
1	5%	1.05	0.02119
2	8%	1.08	0.03342
3	12%	1.12	0.04922
4	7%	1.07	0.02938
5	10%	1.10	0.04139
Total	—	—	0.17460

Applying equation (1.7):

$$log(GM) = \frac{0.17460}{5} = 0.03492,$$

$$GM = antilog(0.03492) = 1.0837.$$

Subtracting 1 to convert back from a growth factor to a growth rate, the average annual growth rate is approximately 8.37%. The simple arithmetic mean of the five percentages would have been (5 + 8 + 12 + 7 + 10)/5 = 8.40%, which is very close but slightly too high. For volatile time series the difference between the two means becomes much larger, and only the geometric mean gives a financially meaningful answer.

Example 1.5 (Geometric Mean — Grouped Data)

The marks of 50 students in a statistics quiz are summarized in the following frequency distribution. Compute the geometric mean of the marks.

Marks (mid-point x)	5	15	25	35	45
Frequency f	4	12	20	10	4

Solution. We extend the table by including columns for log x and f log x.

X	F	log x	f log x
5	4	0.6990	2.7960
15	12	1.1761	14.1132
25	20	1.3979	27.9580
35	10	1.5441	15.4410

45	4	1.6532	6.6128
Total	50	—	66.9210

Applying equation (1.8):

$$log(GM) = \frac{66.9210}{50} = 1.3384,$$

$$GM = antilog(1.3384) \approx 21.80\, marks.$$

The geometric mean of the marks is approximately 21.80. As expected, this value is somewhat smaller than the corresponding arithmetic mean (which the student is invited to verify equals 24.20 marks). It is a general fact that, for any set of positive observations that are not all equal, the geometric mean is always strictly less than the arithmetic mean.

1.6 THE MEDIAN

The median is the value that divides an ordered data set into two equal halves. If we arrange all the observations in order from smallest to largest, the median is the value that lies exactly in the middle of the list, so that fifty per cent of the observations are at or below it and fifty per cent are at or above it. For this reason, the median is sometimes called the 50-th percentile, or equivalently the second quartile, denoted Q_2.

Unlike the arithmetic mean, the median is not affected by extreme observations. A single outlier far above or far below the rest of the data has no effect at all on the median, because the median depends only on the order of the observations and on which one happens to lie in the middle. This robustness makes the median an especially attractive measure of central tendency in the social sciences, where variables such as income, wealth, household expenditure, house prices and reaction times are very often skewed and contain a few unusually large observations. In such situations the median gives a much more honest picture of the "typical" value than the mean does.

1.6.1 Median for Ungrouped Data

Let $x_1, x_2, \ldots, x_n$ be the raw observations. The first step is to arrange them in ascending order of magnitude. The resulting ordered list is called the ordered array, and we denote its values by $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. The median is then defined as follows:

1. If n is odd, the median is the middle observation of the ordered array, namely the value at position $(n + 1)/2$.
2. If n is even, the median is the arithmetic mean of the two middle observations of the ordered array, namely the mean of the values at positions $n/2$ and $n/2 + 1$.

Example 1.6 (Median — Ungrouped Data, Odd and Even Sample Sizes)

(a) Odd sample size. The number of children in nine randomly selected households of a village near Islamabad is 3, 1, 4, 2, 5, 2, 6, 3 and 4. Find the median.

Solution. Arranging the observations in ascending order: 1, 2, 2, 3, 3, 4, 4, 5, 6. Since n = 9 is odd, the median is the value at position (9 + 1)/2 = 5, which is the fifth observation in the ordered array, namely 3. The median number of children per household is therefore 3.

(b) Even sample size. The daily wages (in rupees) of ten labourers employed at a construction site in Islamabad are 800, 750, 950, 700, 900, 850, 1000, 780, 820 and 880. Find the median wage.

Solution. Arranging in ascending order: 700, 750, 780, 800, 820, 850, 880, 900, 950, 1000. Since n = 10 is even, the median is the arithmetic mean of the 5-th and 6-th observations:

$$Median = \frac{820 + 850}{2} = 835\, rupees.$$

Half of the labourers earn no more than Rs. 835 per day, and half earn at least that much.

1.6.2 Median for Grouped Data

When the data are arranged in a frequency distribution, we cannot identify any single observation as the "middle" one, because we no longer know the exact values of the individual observations. Instead, we identify the class in which the middle observation must lie — called the median class — and then use linear interpolation within that class to estimate the median. The standard formula is:

$$Median = l + \frac{h}{f}\left(\frac{n}{2} - F\right) \qquad (1.9)$$

where l is the lower class boundary of the median class, h is the width of the median class, f is the frequency of the median class, n is the total number of observations, and F is the cumulative frequency of the class immediately preceding the median class. The median class is found by computing n/2 and then locating the first class whose cumulative frequency is greater than or equal to n/2.

Example 1.7 (Median — Grouped Data)

Compute the median monthly household expenditure for the data of Example 1.2.

Solution. We first construct the cumulative frequency column.

Class	Class Boundaries	f	Cumulative Frequency (F)
20 – 29	19.5 – 29.5	4	4
30 – 39	29.5 – 39.5	9	13
40 – 49	39.5 – 49.5	14	27
50 – 59	49.5 – 59.5	18	45
60 – 69	59.5 – 69.5	8	53
70 – 79	69.5 – 79.5	5	58
80 – 89	79.5 – 89.5	2	60

Here n/2 = 60/2 = 30. The first cumulative frequency to reach or exceed 30 is 45, which corresponds to the class 50 – 59. So the median class is 50 – 59, with l = 49.5, h = 10, f = 18 and F = 27 (the cumulative frequency of the class just before the median class). Substituting into equation (1.9):

$$Median = 49.5 + \frac{10}{18}(30 - 27) = 49.5 + \frac{30}{18} = 49.5 + 1.67 = 51.17\ thousand\ rupees.$$

The median monthly expenditure is approximately Rs. 51,170. This is almost identical to the mean we computed in Example 1.2, which is a strong indication that the distribution of household expenditures in this sample is fairly symmetric.

1.6.3 Merits and Demerits of the Median

Merits. The median is easy to understand and easy to compute. It is not affected by extreme values, which makes it the natural measure of centre for skewed distributions. It can be computed even when some of the observations at the two extremes are unknown — provided we know how many there are — and it can be used for ordinal qualitative data such as levels of education or severity of illness, for which the arithmetic mean is meaningless.

Demerits. The median is not based on every observation in the data set; changing the values of observations far away from the centre has no effect on the median. It cannot be used for nominal (unordered) qualitative data such as eye colour. And it is less amenable to further algebraic treatment than the mean, which limits its usefulness in advanced inferential techniques.

1.7 THE MODE

The mode is the value that occurs most frequently in a data set. A distribution that has exactly one mode is called unimodal; one with two modes is called bimodal; and one with more than two is called multimodal. When no observation is repeated more often than any other, the data set is said to have no mode. The mode has one big advantage that the mean and the median lack: it is the only one of the three measures of central tendency that can be applied to purely qualitative (categorical) data. We can speak meaningfully of the modal eye colour in a class of students, the modal political party in a constituency, or the modal blood group in a hospital ward — even though it would be nonsense to talk about the mean or the median of these things.

1.7.1 Mode for Ungrouped Data

For ungrouped data, the mode is found simply by inspection: it is the value with the highest frequency in the data set. The student should be careful, however: if every value occurs exactly once, the data set has no mode at all, not "every value as the mode".

Example 1.8 (Mode — Ungrouped Data)

The shoe sizes recorded for fifteen male students of a college in Lahore are 8, 9, 7, 8, 10, 9, 8, 11, 9, 8, 7, 9, 10, 8, 9. Find the mode.

Solution. We tally the frequencies: size 7 occurs 2 times, size 8 occurs 5 times, size 9 occurs 5 times, size 10 occurs 2 times, and size 11 occurs 1 time. Sizes 8 and 9 both occur with the highest frequency (five times each). The data set is therefore bimodal, with modes 8 and 9.

1.7.2 Mode for Grouped Data

For grouped data, the modal class is the class with the highest frequency. Within the modal class, the precise location of the mode is estimated by the formula:

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \cdot h \qquad (1.10)$$

where l is the lower class boundary of the modal class, f_m is the frequency of the modal class, f_1 is the frequency of the class immediately preceding the modal class, f_2 is the frequency of the class immediately following the modal class, and h is the class width. The intuition behind this formula is that, if the class just before the modal class has a much higher frequency than the class just after it, then the mode is pulled towards the left within the modal class; if the situation is reversed, the mode is pulled towards the right.

Example 1.9 (Mode — Grouped Data)

Compute the modal monthly household expenditure for the data of Example 1.2.

Solution. Inspecting the frequency column, the highest frequency is 18, which corresponds to the class 50 – 59. So the modal class has class boundaries 49.5 – 59.5, with l = 49.5, h = 10, f_m = 18, f_1 = 14 (the preceding class) and f_2 = 8 (the following class). Substituting into equation (1.10):

$$Mode = 49.5 + \frac{18 - 14}{(18 - 14) + (18 - 8)} \times 10 = 49.5 + \frac{4}{14} \times 10 = 49.5 + 2.86$$
$$= 52.36 \; thousand \; rupees.$$

The modal monthly expenditure is approximately Rs. 52,360. We now have all three measures of central tendency for this data set: mean = Rs. 51,170, median = Rs. 51,170, mode = Rs. 52,360. The closeness of the three values confirms once again that the distribution of household expenditures is approximately symmetric.

The empirical relationship. For moderately skewed distributions there is a useful empirical relationship, due to Karl Pearson, connecting the three measures of central tendency:

$$Mean - Mode \approx 3\,(Mean - Median) \qquad (1.11)$$

which can be rearranged to give Mode ≈ 3 Median − 2 Mean. This formula is handy when one of the three measures is unknown but the other two are available, particularly when the mode is hard to compute reliably (for example, when the data are continuous and there are no exact ties).

1.8 CHOOSING AN APPROPRIATE MEASURE OF CENTRAL TENDENCY

Now that we have studied five different measures of central tendency — the arithmetic mean, the weighted mean, the geometric mean, the median and the mode — a perfectly natural question arises: which of them should we use in any given situation? There is no single, universal answer to this question, but the following practical guidelines, distilled from Levin and Rubin (2009) and Lind, Marchal and Wathen (2005), should help the student make an informed choice.

1. Use the arithmetic mean when the data are quantitative, the distribution is approximately symmetric, and there are no extreme values. Most inferential techniques in statistics are built upon the mean, so it should be the default choice whenever its use is justified.

2. Use the median when the data are quantitative but the distribution is markedly skewed or contains extreme values. Variables such as income, household expenditure, house prices, waiting times, reaction times and lengths of stay in a hospital are typically best summarized by the median rather than by the mean.
3. Use the mode when the data are qualitative (for example, eye colour, marital status, political party, religion or blood group), or when the most frequently occurring value is itself of direct interest — for instance, the most popular brand of soap, the most common shoe size, or the most frequently chosen examination option.
4. Use the weighted mean whenever the observations differ in importance — for example, in computing a GPA, an index number such as the Consumer Price Index, or the average return on a portfolio.
5. Use the geometric mean whenever the data represent rates of change, ratios, percentages or compounding factors — for example, when computing the average annual growth rate of population, GDP, exports or revenue.

In most applied studies, more than one measure of central tendency is reported in the same table or paragraph, because each measure answers a slightly different question about the "typical" value. A researcher who reports the mean, median and mode together — together with at least one measure of dispersion — gives the reader a much fuller picture of the distribution than one who reports only a single number. We will see this principle at work repeatedly in the chapters that follow.

1.9 QUARTILES, DECILES AND PERCENTILES

The median divides an ordered data set into two equal halves. More generally, we may wish to divide an ordered data set into any number of equal parts. The values that achieve this division are known collectively as quantiles. The most commonly used quantiles are the quartiles, which divide the data into four equal parts; the deciles, which divide it into ten; and the percentiles, which divide it into one hundred. Quantiles are extremely useful in describing the position of an individual observation within a data set: a student who scores at the 90-th percentile of an examination, for example, has done better than 90% of the candidates and worse than only 10% of them.

1.9.1 Quartiles

The three values Q_1, Q_2 and Q_3 divide an ordered data set into four equal parts. Q_1, the first or lower quartile, is the value below which 25% of the observations lie. Q_2, the second quartile, is the median itself. Q_3, the third or upper quartile, is the value below which 75% of the observations lie. The interval from Q_1 to Q_3 therefore contains the central 50% of the data and is called the interquartile range, often abbreviated IQR.

Quartiles for ungrouped data. We first arrange the n observations in ascending order. The position of the j-th quartile in the ordered array is then computed by:

$$Q_j \text{ lies at position } \frac{j(n+1)}{4}, \quad j = 1, 2, 3. \tag{1.12}$$

If the position is not a whole number, linear interpolation is used between the two nearest values of the ordered array. For example, if the position works out to 12.75, then Q is computed as the value at position 12 plus 0.75 times the difference between the values at positions 13 and 12.

Quartiles for grouped data. For data presented as a frequency distribution, the j-th quartile is computed by:

$$Q_j = l + \frac{h}{f}\left(\frac{jn}{4} - F\right), \quad j = 1, 2, 3, \qquad (1.13)$$

where l is the lower class boundary of the j-th quartile class, h is the width of that class, f is its frequency, n is the total sample size, and F is the cumulative frequency up to (but not including) the j-th quartile class. The j-th quartile class is the first class whose cumulative frequency reaches or exceeds $\frac{jn}{4}$.

1.9.2 Deciles and Percentiles

Deciles divide the data into ten equal parts and are denoted $D_1, D_2, \ldots, D_9$. Percentiles divide the data into one hundred equal parts and are denoted $P_1, P_2, \ldots, P_{99}$. Their formulas are direct generalizations of equation (1.13). For grouped data:

$$D_j = l + \frac{h}{f}\left(\frac{jn}{10} - F\right), \quad P_j = l + \frac{h}{f}\left(\frac{jn}{100} - F\right) \qquad (1.14)$$

Note that several of these quantiles are simply different names for the same number: $Q_2 = D_5 = P_{50}$ = Median, $Q_1 = P_{25}$, $Q_3 = P_{75}$, and $D_1 = P_{10}$. The percentile is the most general of the three; quartiles and deciles are simply percentiles with particularly convenient values of j. Percentiles are widely used in education (to report test scores), in medicine (to describe a child's height or weight relative to a reference population), and in economics (to describe the distribution of income).

Example 1.10 (Quartiles — Ungrouped Data)

The marks obtained by twelve students of MA Education in the final examination of "Educational Statistics" at AIOU are 42, 58, 65, 67, 68, 70, 72, 74, 75, 78, 80 and 95. Find Q_1 and Q_3.

Solution. The data are already in ascending order, with n = 12. The position of Q_1 is (n + 1)/4 = 13/4 = 3.25, so Q_1 lies between the 3rd and 4th observations:

$$Q_1 = 65 + 0.25(67 - 65) = 65 + 0.50 = 65.50.$$

The position of Q_3 is 3(n + 1)/4 = 39/4 = 9.75, so Q_3 lies between the 9th and 10th observations:

$$Q_3 = 75 + 0.75(78 - 75) = 75 + 2.25 = 77.25.$$

A quarter of the students scored 65.50 marks or less, and three-quarters scored 77.25 marks or less. The middle half of the class lies between these two values.

Example 1.11 (Quartiles, Deciles and Percentiles — Grouped Data)

For the household expenditure data of Example 1.2, compute Q_1, Q_3, D_7 and P_{90}.

Solution. We use the cumulative frequency table from Example 1.7. Here n = 60, so n/4 = 15, 3n/4 = 45, 7n/10 = 42 and 90n/100 = 54.

Q_1: The first cumulative frequency to reach 15 is 27 (the class 40 – 49). So l = 39.5, h = 10, f = 14, F = 13.

$$Q_1 = 39.5 + \frac{10}{14}(15 - 13) = 39.5 + 1.43 = 40.93 \; thousand\; rupees.$$

Q_3: The first cumulative frequency to reach 45 is 45 itself (the class 50 – 59). So l = 49.5, h = 10, f = 18, F = 27.

$$Q_3 = 49.5 + \frac{10}{18}(45 - 27) = 49.5 + 10.00 = 59.50\ thousand\ rupees.$$

D_7: The first cumulative frequency to reach 42 is 45 (the class 50 – 59). Same class as Q_3.

$$D_7 = 49.5 + \frac{10}{18}(42 - 27) = 49.5 + 8.33 = 57.83\ thousand\ rupees.$$

P_{90}: The first cumulative frequency to reach 54 is 58 (the class 70 – 79). So l = 69.5, h = 10, f = 5, F = 53.

$$P_{90} = 69.5 + \frac{10}{5}(54 - 53) = 69.5 + 2.00 = 71.50\ thousand\ rupees.$$

Interpretation: a quarter of the families in this sample have monthly expenditures below Rs. 40,930; three-quarters have expenditures below Rs. 59,500; seventy per cent are below Rs. 57,830; and only 10% of the families have monthly expenditures greater than Rs. 71,500.

1.10 MEASURES OF DISPERSION: AN OVERVIEW

The measures of central tendency studied so far answer only one of the four questions posed in the introduction: where is the centre of the data? They tell us nothing about the equally important second question: how spread out are the observations around that centre? Two data sets may have identical means, identical medians and identical modes, and yet be dramatically different in character. Consider the following three samples, each of size twelve:

Sample A: 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34

Sample B: 32, 32, 33, 33, 34, 34, 34, 34, 35, 35, 36, 36

Sample C: 21, 24, 26, 28, 31, 34, 34, 37, 39, 42, 45, 47

Each of these three samples has an arithmetic mean of exactly 34, and yet they are obviously very different. Sample A has no variability whatsoever — every observation is precisely equal to the mean. Sample B shows a modest amount of variability, with no observation differing from the mean by more than 2 units. Sample C is clearly the most spread out: some of its observations are as far as 13 units from the mean. To describe a data set adequately, we therefore need a numerical measure of this variability or spread, in addition to a measure of centre. Such measures are called measures of dispersion, and the next several sections introduce the most important of them.

Measures of dispersion are broadly divided into two kinds. Absolute measures of dispersion are expressed in the same units as the original observations and tell us, in those units, how spread out the data are. Examples include the range, the quartile deviation, the mean deviation, the variance and the standard deviation. Relative measures of dispersion, on the other hand, are dimensionless quantities — pure numbers — obtained by dividing an absolute measure by an appropriate measure of central tendency.

Relative measures are used to compare the variability of two or more data sets that are measured in different units, or that have very different means. Examples include the coefficient of range, the coefficient of quartile deviation, the coefficient of mean deviation and the coefficient of variation. Each of the absolute measures presented below has its own corresponding relative form, as we shall see.

Measures of dispersion are broadly divided into two kinds. Absolute measures of dispersion are expressed in the same units as the original observations and tell us, in those units, how spread out the data are. Examples include the range, the quartile deviation, the mean deviation, the variance and the standard deviation. Relative measures of dispersion, on the other hand, are dimensionless quantities — pure numbers — obtained by dividing an absolute measure by an appropriate measure of central tendency. Relative measures are used to compare the variability of two or more data sets that are measured in different units, or that have very different means. Examples include the coefficient of range, the coefficient of quartile deviation, the coefficient of mean deviation and the coefficient of variation.

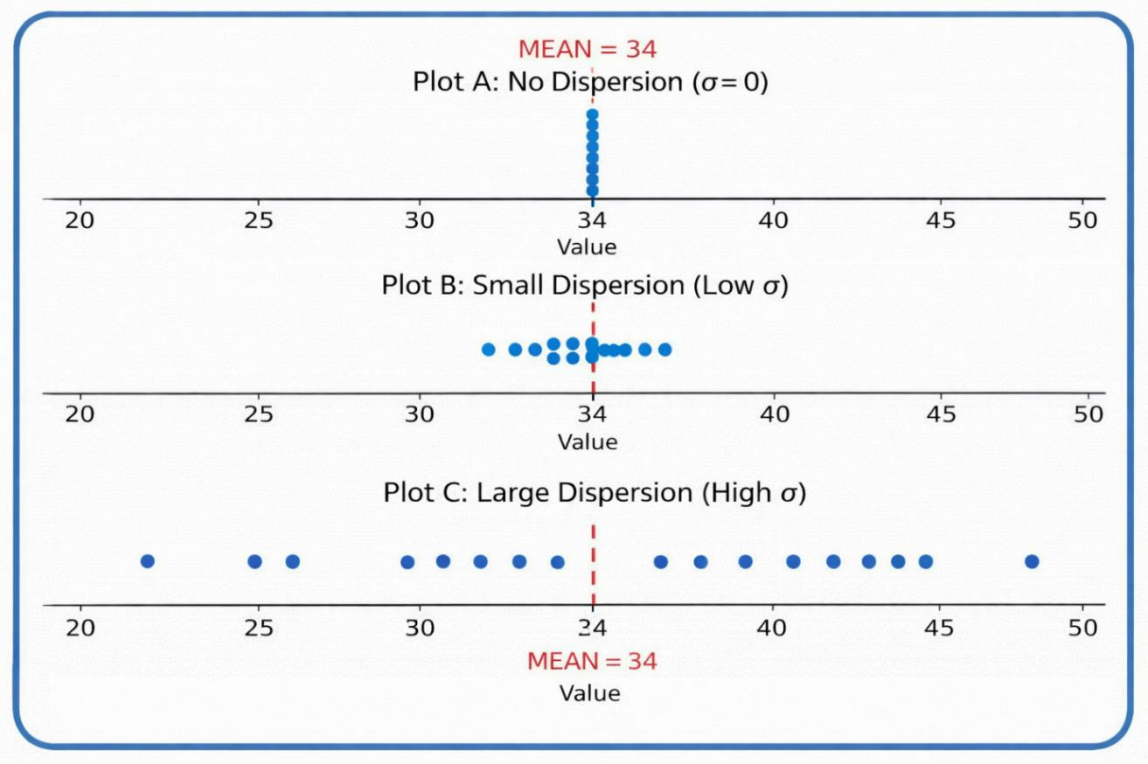


Figure 1.1 Why the Mean Alone is Insufficient? A Visualization of Dispersion

1.11 THE RANGE

The range R is the simplest of all measures of dispersion. It is defined as the difference between the largest and the smallest observations in the data set. The range gives a quick rough idea of the total spread of the data, and for that reason it is widely used in industrial quality control and in everyday weather reports (in which the daily minimum and maximum temperatures are reported). Its drawbacks are equally well known, however: the range is based on only two of the observations, and so it ignores almost all the information contained in the data; and it is highly sensitive to outliers, because a single unusually large or small value can inflate the range enormously.

1.11.1 Range for Ungrouped Data

For ungrouped data with smallest value $x_{(1)}$ and largest value $x_{(n)}$ from the ordered array, the range and the corresponding coefficient of range are defined by:

$$R = x_{(n)} - x_{(1)} \quad and \quad Coefficient\ of\ Range = \frac{x_{(n)} - x_{(1)}}{x_{(n)} + x_{(1)}} \tag{1.15}$$

1.11.2 Range for Grouped Data

For data presented in a frequency distribution, the range is taken as the difference between the upper class boundary of the highest class and the lower class boundary of the lowest class:

$$R = x_{uh} - x_{ll} \quad and \quad Coefficient\ of\ Range = \frac{x_{uh} - x_{ll}}{x_{uh} + x_{ll}} \qquad (1.16)$$

where x_{ll} is the lower class boundary of the lowest class and xuh is the upper class boundary of the highest class.

Example 1.12 (Range — Ungrouped Data)

The heights (in centimetres) of ten second-semester students of BS Statistics measured to the nearest whole number are 156, 171, 162, 165, 159, 167, 164, 168, 170 and 163. Compute the range and the coefficient of range.

Solution. The smallest value is $x_{(1)}$ = 156 cm and the largest is $x_{(n)}$ = 171 cm. Therefore:

$$R = 171 - 156 = 15\ cm,$$

$$Coefficient\ of\ Range = \frac{171 - 156}{171 + 156} = \frac{15}{327} = 0.0459 = 4.59\%.$$

The total spread of the heights is 15 cm, which represents about 4.59% of the sum of the smallest and largest heights.

Example 1.13 (Range — Grouped Data)

Compute the range and the coefficient of range for the household expenditure data of Example 1.2.

Solution. The lower class boundary of the lowest class (20 – 29) is xll = 19.5, and the upper class boundary of the highest class (80 – 89) is xuh = 89.5. Therefore:

$$R = 89.5 - 19.5 = 70\ thousand\ rupees,$$

$$Coefficient\ of\ Range = \frac{89.5 - 19.5}{89.5 + 19.5} = \frac{70}{109} = 0.6422 = 64.22\%.$$

The total spread of monthly expenditures is Rs. 70,000, which equals about 64.22% of the sum of the two extreme boundary values. The coefficient of range is much larger here than in the previous example because the household expenditure data are inherently far more variable, in relative terms, than the height data.

1.12 THE QUARTILE DEVIATION (SEMI-INTERQUARTILE RANGE)

A measure of spread that is much less affected by outliers than the range is obtained by using the quartiles. The interquartile range, IQR = $Q_3 - Q_1$, measures the spread of the middle 50% of the observations. Half of this quantity — the semi-interquartile range — is called the quartile deviation, denoted QD. By using only the central half of the data, the quartile deviation completely ignores the extreme observations at the two ends of the distribution and is therefore unaffected by outliers. It can also be computed for distributions with open-ended class intervals at one or both extremes (for example, "less than 10" or "80 and above"), where the range cannot be computed at all.

1.12.1 Quartile Deviation for Ungrouped Data

The quartile deviation and its corresponding relative measure, the coefficient of quartile deviation, are defined by:

$$QD = \frac{Q_3 - Q_1}{2} \quad and \quad Coefficient\ of\ QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{1.17}$$

Q_1 and Q_3 are computed from the ordered array as in Section 1.9.1, using equation (1.12).

1.12.2 Quartile Deviation for Grouped Data

For data presented in a frequency distribution, Q_1 and Q_3 are computed from equation (1.13) and the quartile deviation is then obtained from equation (1.17). The formulas for QD and the coefficient of QD are formally identical to the ungrouped case; only the way of computing Q_1 and Q_3 changes.

Example 1.14 (Quartile Deviation — Ungrouped Data)

Using the marks data of Example 1.10, compute the quartile deviation and the coefficient of quartile deviation.

Solution. From Example 1.10 we already have $Q_1 = 65.50$ and $Q_3 = 77.25$. Substituting into equation (1.17):

$$QD = \frac{77.25 - 65.50}{2} = \frac{11.75}{2} = 5.875\ marks.$$

$$Coefficient\ of\ QD = \frac{77.25 - 65.50}{77.25 + 65.50} = \frac{11.75}{142.75} = 0.0823 = 8.23\%.$$

The middle half of the marks is contained within an interval of width $2 \times 5.875 = 11.75$ marks, centred (roughly) on the median.

Example 1.15 (Quartile Deviation — Grouped Data)

Using the household expenditure data of Example 1.2 and the quartiles computed in Example 1.11 ($Q_1 = 40.93$ and $Q_3 = 59.50$ thousand rupees), compute the quartile deviation and the coefficient of quartile deviation.

Solution. Substituting directly into equation (1.17):

$$QD = \frac{59.50 - 40.93}{2} = \frac{18.57}{2} = 9.285\ thousand\ rupees.$$

$$Coefficient\ of\ QD = \frac{59.50 - 40.93}{59.50 + 40.93} = \frac{18.57}{100.43} = 0.1849 = 18.49\%.$$

The middle half of the families have monthly expenditures lying within an interval of width Rs. 18,570 around the median. Notice that the coefficient of quartile deviation (18.49%) is much smaller than the coefficient of range (64.22%) we computed in Example 1.13. The reason is that the quartile deviation completely ignores the small number of families at the two extremes, while the range is dominated by them. This dramatic numerical difference illustrates very clearly why the quartile deviation is preferred to the range whenever robustness against outliers is important.

1.13 THE MEAN DEVIATION

A natural way of measuring how spread out a data set is would be to compute the deviations of all the observations from some central value (such as the mean or the median) and then to take the average of those deviations. Unfortunately, as we noted in Section 1.3.3, the sum of the deviations from the mean is always exactly zero, so a simple average of these deviations is useless as a measure of spread. There are two standard ways of getting around this difficulty. One is to take the average of the absolute values of the deviations, ignoring the signs; this gives the mean deviation, which we study in this section. The other is to take the average of the squared deviations; this gives the variance and the standard deviation, which we study in the next section. The mean deviation is more intuitive and easier to interpret than the standard deviation, but the standard deviation has much better mathematical properties, which is why it has come to dominate inferential statistics.

The mean deviation, also called the mean absolute deviation, is defined as the arithmetic mean of the absolute deviations of the observations from some central value, normally either the arithmetic mean or the median. It can be shown mathematically that the sum of the absolute deviations from the median is smaller than from any other value, so the mean deviation about the median is always less than or equal to the mean deviation about the mean. For this reason some textbooks recommend computing the mean deviation about the median rather than about the mean.

1.13.1 Mean Deviation for Ungrouped Data

For a sample of n observations $x_1, x_2, \dots, x_n$ with sample mean $\bar{x}$ and sample median $\tilde{x}$, the mean deviation about the mean and about the median are defined respectively by:

$$MD(mean) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad , \quad MD(median) = \frac{\sum_{i=1}^{n} |x_i - \tilde{x}|}{n} \qquad (1.18)$$

The corresponding relative measures, called the coefficients of mean deviation, are obtained by dividing each mean deviation by the central value from which the deviations were taken:

$$Coefficient\ of\ MD = \frac{MD(mean)}{\bar{x}} \quad or \quad \frac{MD(median)}{\tilde{x}} \qquad (1.19)$$

1.13.2 Mean Deviation for Grouped Data

For a frequency distribution with class mid-points $x_1, x_2, \dots, x_k$ and frequencies $f^1, f^2, \dots, f_k$ (with $\Sigma f_i = n$), the mean deviations are computed by:

$$MD(mean) = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{x}|}{\sum_{i=1}^{k} f_i} \qquad (1.20)$$

and similarly with $\tilde{x}$ in place of $\bar{x}$ for the mean deviation about the median. Note that for any data set the standard deviation is always at least as large as the mean deviation, because the squaring operation in the standard deviation gives more weight to large deviations than the absolute-value operation in the mean deviation does.

Example 1.16 (Mean Deviation — Ungrouped Data)

The weights (in kilograms) of ten second-semester students of BS Statistics, measured to one decimal place, are 37.7, 40.3, 43.3, 44.5, 46.9, 47.6, 48.6, 51.5, 52.4 and 53.8. Compute the mean deviation about the mean and the corresponding coefficient.

Solution. We first compute the mean:

$$\bar{x} = \frac{37.7 + 40.3 + 43.3 + 44.5 + 46.9 + 47.6 + 48.6 + 51.5 + 52.4 + 53.8}{10}$$
$$= \frac{466.6}{10} = 46.66\ kg.$$

Next we compute the absolute deviations from the mean:

x	$\lvert x - \bar{x} \rvert$
37.7	8.96
40.3	6.36
43.3	3.36
44.5	2.16
46.9	0.24
47.6	0.94
48.6	1.94
51.5	4.84
52.4	5.74
53.8	7.14
Total	41.68

Applying equation (1.18):

$$MD(mean) = \frac{41.68}{10} = 4.17\ kg.$$

$$Coefficient\ of\ MD = \frac{4.17}{46.66} = 0.0894 = 8.94\%.$$

On average, an individual weight in this sample differs from the mean by about 4.17 kg, or about 8.94% of the mean.

Example 1.17 (Mean Deviation — Grouped Data)

Compute the mean deviation about the mean for the household expenditure data of Example 1.2.

Solution. From Example 1.2 we have $\bar{x} = 51.17$ thousand rupees. We extend the calculation table by adding columns for $|x - \bar{x}|$ and $f|x - \bar{x}|$.

Class	f	x	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
20 – 29	4	24.5	26.67	106.68
30 – 39	9	34.5	16.67	150.03
40 – 49	14	44.5	6.67	93.38

50 – 59	18	54.5	3.33	59.94
60 – 69	8	64.5	13.33	106.64
70 – 79	5	74.5	23.33	116.65
80 – 89	2	84.5	33.33	66.66
Total	60	—	—	699.98

Applying equation (1.20):

$$MD(mean) = \frac{699.98}{60} = 11.67\ thousand\ rupees.$$

$$Coefficient\ of\ MD = \frac{11.67}{51.17} = 0.2281 = 22.81\%.$$

On average, an individual family's monthly expenditure differs from the mean of Rs. 51,170 by about Rs. 11,670, or about 22.81% of the mean. As we shall see in the next section, this is somewhat smaller than the standard deviation of the same data set, which is the general rule.

1.14 THE VARIANCE AND THE STANDARD DEVIATION

The variance and the standard deviation are by far the most important measures of dispersion in statistical theory and practice. They are based on the idea of squaring the deviations from the mean before averaging, which automatically takes care of the sign problem (since a square is never negative) and at the same time gives extra weight to large deviations. The squared units of the variance are sometimes inconvenient — a variance of 169 "square rupees" is hard to interpret directly — and so we usually take the positive square root to obtain the standard deviation, which has the same units as the original observations and is therefore easier to interpret.

1.14.1 Variance and Standard Deviation — Definitions

For a population of N observations $X_1, X_2, \ldots, X_N$ with population mean μ, the population variance σ^2 and the population standard deviation σ are defined by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \quad , \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \qquad (1.21)$$

For a sample of n observations $x_1, x_2, \ldots, x_n$ drawn from such a population, with sample mean $\bar{x}$, the sample variance s^2 and the sample standard deviation s are defined by:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad , \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (1.22)$$

A critical observation about the divisor. Notice carefully that the population variance divides the sum of squared deviations by N, whereas the sample variance divides by (n − 1), not by n. The divisor (n − 1) is called the number of degrees of freedom of the sample variance. The reason for using (n − 1) rather than n in the sample formula is that doing so makes s^2 an unbiased estimator of the population variance σ^2. If we divided by n instead, we would consistently underestimate σ^2. The unbiasedness of s^2 will be proved properly in Unit 3 on Estimation; for the moment the student should simply remember the rule: divide by N for a population variance and by (n − 1) for a sample variance. For very large samples (say, n greater than about

100) the practical difference between dividing by n and dividing by (n − 1) is negligible; for small samples it can be substantial.

Some older textbooks, including some editions of Lind, Marchal and Wathen, present a "sample variance" that divides by n. Such a quantity is sometimes called the biased sample variance. In this course, in line with Levin and Rubin (2009), we adopt the unbiased version with the divisor (n − 1) for sample data, and the divisor N for population data, throughout.

1.14.2 Computing Formulas

The definitional formulas (1.21) and (1.22) require us first to compute the mean and then to calculate every individual deviation, which is rather tedious for large data sets. An algebraically equivalent computing formula avoids this by expressing the sum of squared deviations directly in terms of the raw observations and their squares:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (1.23)$$

so that the sample variance can be computed as:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right) \qquad (1.24)$$

1.14.3 Variance and Standard Deviation for Grouped Data

For data presented in a frequency distribution with class mid-points $x_1, x_2, \ldots, x_k$ and frequencies $f_1, f_2, \ldots, f_k$ (with $\Sigma f_i = n$), the sample variance is computed by:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1} \qquad (1.25)$$

with the corresponding computing formula:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}\right) \qquad (1.26)$$

A short-cut coding method, exactly analogous to the one we used for the mean in Section 1.3.2, is also available. Using $u_i = (x_i - a)/h$ with provisional mean a and class width h, the variance becomes:

$$s^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{k} f_i u_i^2 - \frac{\left(\sum_{i=1}^{k} f_i u_i\right)^2}{n}\right) \qquad (1.27)$$

Example 1.18 (Variance and Standard Deviation — Ungrouped Data)

The following data represent the ages (in years) of eight employees of a small marketing firm in Islamabad: 28, 32, 35, 30, 42, 38, 27, 34. Compute the sample mean, the sample variance and the sample standard deviation, using both the definitional formula (1.22) and the computing formula (1.24).

Solution. We prepare a calculation table.

x	x − x̄	(x − x̄)²	x²
28	−5.25	27.5625	784
32	−1.25	1.5625	1024
35	1.75	3.0625	1225

30	−3.25	10.5625	900
42	8.75	76.5625	1764
38	4.75	22.5625	1444
27	−6.25	39.0625	729
34	0.75	0.5625	1156
Total = 266	0.00	181.5000	9026

The sample mean is:

$$\bar{x} = \frac{266}{8} = 33.25\ years.$$

(a) Definitional formula (1.22):

$$s^2 = \frac{181.50}{8 - 1} = \frac{181.50}{7} = 25.93, \quad s = \sqrt{25.93} = 5.09\ years.$$

(b) Computing formula (1.24):

$$s^2 = \frac{1}{7}\left(9026 - \frac{(266)^2}{8}\right) = \frac{1}{7}(9026 - 8844.5) = \frac{181.5}{7} = 25.93.$$

Both methods agree exactly. The sample standard deviation of employee ages is approximately 5.09 years, indicating that most employees are within about 5 years of the mean age of 33.25.

Example 1.19 (Variance and Standard Deviation — Grouped Data)

Compute the variance and standard deviation for the household expenditure data of Example 1.2 using both the direct method (equation 1.25) and the coding method (equation 1.27).

Solution. We extend the calculation table from Example 1.2 by adding columns for $(x - \bar{x})^2$, $f(x - \bar{x})^2$, u, fu and fu^2. Recall that $\bar{x} = 51.17$, a = 54.5 and h = 10.

Class	f	X	$x - \bar{x}$	$f(x - \bar{x})^2$	u	Fu	fu^2
20 – 29	4	24.5	−26.67	2845.38	−3	−12	36
30 – 39	9	34.5	−16.67	2500.44	−2	−18	36
40 – 49	14	44.5	−6.67	622.96	−1	−14	14
50 – 59	18	54.5	3.33	199.56	0	0	0
60 – 69	8	64.5	13.33	1421.42	1	8	8
70 – 79	5	74.5	23.33	2721.65	2	10	20
80 – 89	2	84.5	33.33	2221.38	3	6	18
Total	60	—	—	12532.79	—	−20	132

(a) Direct method (equation 1.25):

$$s^2 = \frac{12532.79}{60 - 1} = \frac{12532.79}{59} = 212.42,$$

$$s = \sqrt{212.42} = 14.58\ thousand\ rupees.$$

(b) Coding method (equation 1.27) with a = 54.5 and h = 10:

$$s^2 = \frac{100}{59}\left(132 - \frac{(-20)^2}{60}\right) = \frac{100}{59}(132 - 6.67) = \frac{12533}{59} = 212.42.$$

$$s = \sqrt{212.42} = 14.58\ thousand\ rupees.$$

Both methods give the same answer. The standard deviation of about Rs. 14,580 tells us that the monthly expenditures of most families fall within about Rs. 14,580 of the mean of Rs. 51,170. Notice also that this standard deviation is larger than the mean deviation of Rs. 11,670 we computed in Example 1.17 — exactly as the theory predicts.

1.15 THE COEFFICIENT OF VARIATION

The standard deviation is expressed in the same units as the original data: rupees, kilograms, minutes, and so on. This is both a strength and a weakness. It is a strength because the interpretation of the standard deviation is direct — we can say "the standard deviation is fourteen thousand rupees" and the listener immediately knows what we mean. It is a weakness because two standard deviations measured in different units cannot be directly compared. We cannot meaningfully say whether the variability of student heights (in centimetres) is greater than the variability of student weights (in kilograms), because the two are measured in completely different units. Even when both variables are measured in the same units, a standard deviation of Rs. 100 represents a great deal of variability for the monthly pocket money of school children but almost no variability for the monthly expenditures of large households.

The solution to this problem is to express the standard deviation as a percentage of the mean. The resulting dimensionless quantity is called the coefficient of variation, abbreviated CV, and is defined for a sample and a population by:

$$CV = \frac{s}{\bar{x}} \times 100\% \quad (sample) \qquad CV = \frac{\sigma}{\mu} \times 100\% \quad (population) \tag{1.28}$$

A higher CV indicates a greater degree of relative variability. When two or more data sets are compared, the set with the smaller CV is regarded as more consistent (less variable in relative terms), and the set with the larger CV as less consistent (more variable in relative terms). The CV is widely used in finance, biology and quality control, where comparisons across different units or different scales are routine.

Example 1.20 (Coefficient of Variation — Comparing Two Groups)

The monthly salaries of employees in two departments of a software house in Islamabad are summarized as follows:

	Department A	Department B
Mean	Rs. 85,000	Rs. 50,000
Standard Deviation	Rs. 12,000	Rs. 8,000

Which department has the more consistent salary structure?

Solution. A direct comparison of the two standard deviations would be misleading, because the two departments have very different mean salaries. We compute the coefficient of variation for each department:

$$CV_A = \frac{12000}{85000} \times 100\% = 14.12\%,$$

$$CV_B = \frac{8000}{50000} \times 100\% = 16.00\%.$$

Although Department A has the larger absolute standard deviation, its coefficient of variation is the smaller of the two (14.12% versus 16.00%). Department A is therefore the more consistent in its salary structure. This example illustrates very clearly why the CV is preferred over the raw standard deviation whenever we are comparing the variability of two or more groups with different mean values.

1.16 INTERPRETING THE STANDARD DEVIATION: CHEBYSHEV AND THE EMPIRICAL RULE

Having introduced several measures of dispersion, we now pause and ask: what does it really mean for a standard deviation to be "large" or "small"? How can we attach a concrete interpretation to such numbers? Two well-known rules — Chebyshev's inequality and the empirical rule — provide practical answers.

1.16.1 Chebyshev's Inequality (for Any Distribution)

Chebyshev's inequality, named after the Russian mathematician P. L. Chebyshev, states that for any data set, and for any constant k strictly greater than 1, at least $(1 - 1/k^2)$ of the observations must lie within k standard deviations of the mean. In symbols, for a population:

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \qquad (1.29)$$

Chebyshev's inequality is remarkable for its generality: it applies to any distribution whatsoever, no matter how skewed or strangely shaped. For example, setting k = 2, at least $1 - 1/4 = 75\%$ of the observations in any data set must lie within two standard deviations of the mean. For k = 3, at least $1 - 1/9 \approx 88.9\%$ of the observations must lie within three standard deviations of the mean. Because Chebyshev's inequality is distribution-free, the bound it provides is not usually very tight, but it is always valid and gives the worst-case guarantee.

1.16.2 The Empirical Rule (for Bell-Shaped Distributions)

When a distribution is approximately bell-shaped — that is, approximately normal, an idea we shall study in detail in Unit 2 — a much sharper rule, known as the empirical rule or the 68-95-99.7 rule, applies:

1. Approximately 68% of the observations lie within one standard deviation of the mean: in the interval $(\bar{x} - s, \bar{x} + s)$.
2. Approximately 95% of the observations lie within two standard deviations of the mean: in the interval $(\bar{x} - 2s, \bar{x} + 2s)$.
3. Approximately 99.7% of the observations lie within three standard deviations of the mean: in the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

The empirical rule is very useful in practice. For example, if the monthly household expenditures in a town are approximately normally distributed with mean Rs. 51,170 and standard deviation Rs. 14,580, then

roughly 68% of households have expenditures between Rs. 36,590 and Rs. 65,750, roughly 95% have expenditures between Rs. 22,010 and Rs. 80,330, and almost all households (99.7%) have expenditures between Rs. 7,430 and Rs. 94,910. Notice the interesting fact that, for bell-shaped data, almost no observations lie more than three standard deviations away from the mean — which is one reason why values lying further than 3s from the mean are often flagged as possible outliers.

1.17 SKEWNESS AND KURTOSIS

Even after we have computed a measure of central tendency and a measure of dispersion, our description of a data set is not yet complete. Two distributions can have identical means and identical standard deviations and still differ dramatically in their overall shape. In particular, they may differ in two further respects: how symmetric they are about the mean, and how sharply they peak. The first of these two characteristics is measured by skewness, and the second by kurtosis. Together with the mean and the standard deviation, the measures of skewness and kurtosis enable us to write a complete numerical summary of any distribution we encounter.

1.17.1 Skewness

A distribution is said to be symmetric if its left and right halves are mirror images of each other when reflected about the mean. For a perfectly symmetric distribution, the mean, the median and the mode are all equal. A distribution that is not symmetric is called skewed. Skewness can be either positive (also called right-skewed) or negative (also called left-skewed).

1. A positively skewed (right-skewed) distribution has a long tail extending to the right of the centre. The bulk of the observations are concentrated towards the left, with a few unusually large observations stretching the right-hand tail. In a positively skewed distribution, the mean is greater than the median, which is in turn greater than the mode. Examples of positively skewed variables include household income, household wealth, house prices, waiting times for public services, and lengths of stay in hospital — in each case, most people are at the modest end and a few are at the very high end.
2. A negatively skewed (left-skewed) distribution has a long tail extending to the left. The bulk of the observations are concentrated towards the right, with a few unusually small observations stretching the left-hand tail. In a negatively skewed distribution, the mean is less than the median, which is in turn less than the mode. Examples include examination scores on a very easy test (most students do well, but a few do badly), the age at death in a developed country (most people die old, but a few die young), and the time taken by experienced workers to perform a routine task.

1.17.2 Kurtosis

Kurtosis is a measure of the "peakedness" of a distribution — how sharply it is peaked at the centre and how heavy its tails are. Three standard categories are recognized:

1. A distribution that is more sharply peaked than the normal distribution, with heavier tails, is called leptokurtic. Such distributions have many observations very close to the mean and also a few that are very far from the mean.
2. A distribution that has the same peakedness as the normal distribution is called mesokurtic. The normal distribution is the standard reference mesokurtic distribution.
3. A distribution that is flatter than the normal distribution, with lighter tails and observations spread more uniformly, is called platykurtic.

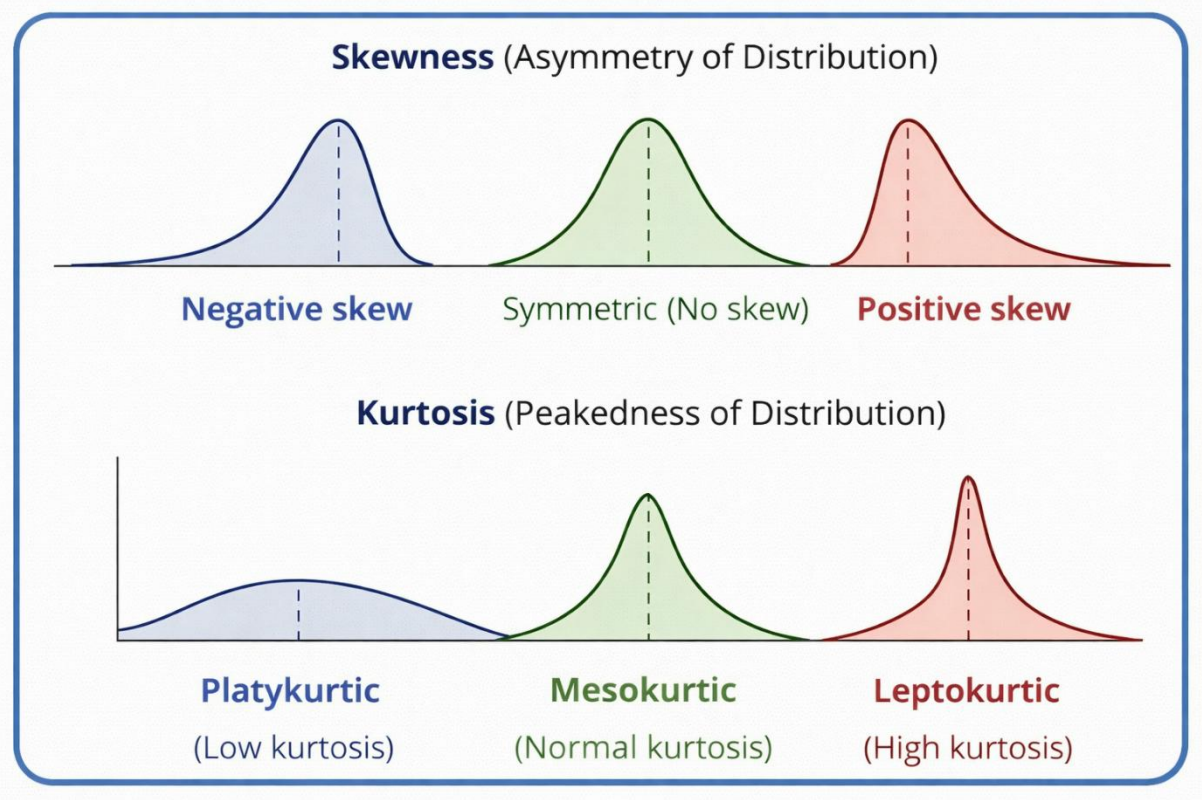


Figure 1.2 Skewness vs Kurtosis

1.18 MEASURES OF SKEWNESS AND PEAKEDNESS

Several different numerical measures of skewness and kurtosis are in common use. We present here the three most important measures of skewness and the standard moment-based measure of kurtosis. Each measure has its own advantages: Karl Pearson's coefficient is intuitive and easy to compute; Bowley's coefficient is robust against outliers; and the moment coefficient is the one most often reported by statistical software.

1.18.1 Karl Pearson's Coefficient of Skewness

Karl Pearson proposed a simple measure of skewness based on the difference between the mean and the mode, divided by the standard deviation:

$$Sk_P = \frac{Mean - Mode}{Standard\ Deviation} \qquad (1.30)$$

When the mode is not well defined — for instance, when the distribution is bimodal or when the raw data have been smoothed — Pearson proposed an alternative version of the formula based on the empirical relationship (1.11) between the mean, median and mode:

$$Sk_P = \frac{3(Mean - Median)}{Standard\ Deviation} \qquad (1.31)$$

Pearson's coefficient is normally between −3 and +3 (although values close to those extremes are rare in practice). A value of zero indicates perfect symmetry; a positive value indicates positive skewness; and a negative value indicates negative skewness. The further the coefficient is from zero, the greater the degree of skewness.

1.18.2 Bowley's Coefficient of Skewness

When the data contain extreme values, or when the distribution has open-ended class intervals at one or both extremes, a quartile-based measure of skewness is preferable. Bowley's coefficient of skewness is defined as:

$$Sk_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \qquad (1.32)$$

Bowley's coefficient always lies between −1 and +1. It is equal to zero for a symmetric distribution (because Q_1 and Q_3 are then equally spaced from the median Q_2), positive for a right-skewed distribution, and negative for a left-skewed distribution. Because it is built entirely from the three quartiles, Bowley's coefficient is unaffected by outliers and can be computed even for distributions with open-ended classes.

1.18.3 The Moment Coefficient of Skewness and Kurtosis

The most widely used theoretical measures of skewness and kurtosis are based on the central moments of the distribution. Define the r-th central moment of the sample as:

$$m_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n} \qquad (1.33)$$

The second central moment m_2 is essentially the population variance (with the divisor n rather than n − 1). The third central moment m_3 measures asymmetry and the fourth central moment m_4 measures peakedness. The moment coefficient of skewness, often denoted by $\sqrt{b_1}$, and the moment coefficient of kurtosis, often denoted by b_2, are defined by:

$$\sqrt{b_1} = \frac{m_3}{s^3} \quad and \quad b_2 = \frac{m_4}{s^4} \qquad (1.34)$$

The interpretation of the moment skewness coefficient $\sqrt{b_1}$ is the same as for Pearson's coefficient: zero means symmetric, positive means right-skewed, and negative means left-skewed. The interpretation of the moment kurtosis coefficient b_2 is as follows: for a normal distribution, $b_2 = 3$. A value of b_2 greater than 3 indicates a leptokurtic distribution (more peaked than normal); a value less than 3 indicates a platykurtic distribution (flatter than normal); and a value equal to 3 indicates a mesokurtic distribution (the same peakedness as normal). Most statistical software packages report what is called "excess kurtosis", defined as $(b_2 - 3)$, so that a value of zero corresponds to the normal distribution, positive values to leptokurtic, and negative values to platykurtic distributions.

Example 1.21 (Measures of Skewness — Ungrouped Data)

For the ten student weights of Example 1.16 (37.7, 40.3, 43.3, 44.5, 46.9, 47.6, 48.6, 51.5, 52.4, 53.8), compute Karl Pearson's second coefficient of skewness and comment on the shape of the distribution.

Solution. We have already computed $\bar{x} = 46.66$ from Example 1.16. The median (the average of the 5th and 6th observations) is (46.9 + 47.6)/2 = 47.25. We also need the standard deviation. The deviations and their squares are:

X	$x - \bar{x}$	$(x - \bar{x})^2$
37.7	−8.96	80.2816
40.3	−6.36	40.4496
43.3	−3.36	11.2896
44.5	−2.16	4.6656
46.9	0.24	0.0576

47.6	0.94	0.8836
48.6	1.94	3.7636
51.5	4.84	23.4256
52.4	5.74	32.9476
53.8	7.14	50.9796
Total	0.00	248.7440

$$s^2 = \frac{248.7440}{10 - 1} = 27.638, \quad s = 5.257\ kg.$$

Applying Pearson's second formula (1.31):

$$Sk_P = \frac{3(46.66 - 47.25)}{5.257} = \frac{-1.77}{5.257} = -0.337.$$

The coefficient is small in absolute value and slightly negative, indicating that the distribution of weights is approximately symmetric with a slight tendency towards negative skewness — a few of the lighter students pull the left tail outwards just a little.

Example 1.22 (Measures of Skewness — Grouped Data)

For the household expenditure data of Example 1.2, we already have the following summary values: mean $\bar{x}$ = 51.17, median = 51.17, mode = 52.36, Q_1 = 40.93, Q_2 = 51.17, Q_3 = 59.50 and standard deviation s = 14.58 (all in thousand rupees). Compute Karl Pearson's and Bowley's coefficients of skewness, and comment on the shape of the distribution.

Solution.

Pearson's first formula (1.30):

$$Sk_P = \frac{51.17 - 52.36}{14.58} = \frac{-1.19}{14.58} = -0.082.$$

Bowley's formula (1.32):

$$Sk_B = \frac{59.50 + 40.93 - 2(51.17)}{59.50 - 40.93} = \frac{-1.91}{18.57} = -0.103.$$

Both coefficients are very small in absolute value and both are slightly negative. We conclude that the distribution of household expenditures in this sample is very nearly symmetric with a small tendency towards negative skewness. This conclusion is consistent with our earlier observation that the mean and the median are practically identical and that the mode is only slightly larger.

1.19 SUMMARY

In this unit we have developed the basic tools of descriptive statistics, which together allow a social-science researcher to summarize a raw data set in a compact and informative way. The main points may be recapitulated as follows.

- A measure of central tendency is a single value that describes the centre of a data set. The three most important measures are the arithmetic mean, the median and the mode. The arithmetic mean is based on every observation and has the most useful mathematical properties, but it is sensitive to extreme values. The median is the middle observation of the ordered data and is unaffected by outliers, which makes it the natural measure of centre for skewed distributions. The mode is the most frequent observation and is the only one of the three that can be applied to purely qualitative data. Two specialized variants — the weighted mean and the geometric mean — are used for observations of unequal importance and for rates of change respectively.
- Quantiles describe the position of individual observations within a data set. Quartiles divide the data into four equal parts; deciles into ten; percentiles into one hundred. The first and third quartiles, Q_1 and Q_3, bracket the middle 50% of the observations and are the basis of several robust measures of spread and skewness.
- A measure of dispersion describes how spread out the observations are around their centre. Simple measures such as the range and the quartile deviation are easy to compute but use very little of the information available. The mean deviation, the variance and the standard deviation use every observation. The sample variance divides the sum of squared deviations by $(n - 1)$ rather than n in order to produce an unbiased estimate of the population variance. The coefficient of variation expresses the standard deviation as a percentage of the mean and is the appropriate tool for comparing the variability of data sets with different means or different units.
- Chebyshev's inequality and the empirical rule provide concrete interpretations of the standard deviation. Chebyshev's inequality holds for any distribution; the empirical 68-95-99.7 rule is sharper but applies only to bell-shaped distributions.
- Skewness measures the asymmetry of a distribution and kurtosis measures its peakedness. Karl Pearson's coefficient and Bowley's coefficient give simple measures of skewness; the moment-based coefficients $\sqrt{b_1}$ and b_2 are the standard theoretical measures used by software packages. Positive skewness, such as that typically found in income and expenditure distributions, signals a long right tail and is associated with the inequality Mean > Median > Mode.

The descriptive tools developed in this unit are not an end in themselves: they form the essential first step in any statistical investigation. In the units that follow, we shall use them again and again as we build up the more sophisticated techniques of statistical inference — confidence intervals, hypothesis tests, regression and correlation analysis, analysis of variance, and forecasting from time series. A firm grasp of the material of Unit 1 is therefore indispensable for everything that follows in this course.

1.20 SELF-ASSESSMENT EXERCISE

1. Define a population and a sample. Using suitable notation, distinguish clearly between a population parameter and a sample statistic. Why is it customary in statistics to use Greek letters for population quantities and Latin letters for sample quantities?
2. State the principal merits and demerits of the arithmetic mean. Under what circumstances would you prefer the median or the mode to the mean as a measure of central tendency?
3. Distinguish between absolute and relative measures of dispersion. Give two examples of each kind, and explain when each is most appropriate.

4. Why do we divide by $(n - 1)$ rather than by n when computing the sample variance? Briefly explain the concept of degrees of freedom as it applies to this formula.

5. Define skewness and kurtosis. With the help of rough sketches, describe the three types of skewness and the three types of kurtosis that a distribution may exhibit.

6. Discuss in detail the arithmetic mean, the weighted mean and the geometric mean. For each of these three measures, present the definition, the formulas for ungrouped and grouped data, at least one practical situation in which the measure is most appropriate, and its main properties. Compare and contrast the three in terms of their strengths and limitations.

7. Describe the variance, the standard deviation and the coefficient of variation as measures of dispersion. Derive the computing formula for the sample variance from the definitional formula. State Chebyshev's inequality and the empirical rule, and explain how each gives a concrete practical interpretation of the standard deviation. Illustrate your discussion with an example of your own choosing.

8. The weights (in kilograms) of twelve first-year students of BS Psychology at AIOU are 58, 62, 65, 60, 72, 68, 55, 70, 63, 67, 59, 64. Compute the sample mean, median, mode and sample standard deviation. [Answers: $\bar{x} \approx 63.58$ kg; Median = 63.5 kg; No unique mode; $s \approx 5.10$ kg]

9. A survey of the daily expenditure on travel (in rupees) of 50 students of Islamabad gave the following frequency distribution:

Expenditure	**50 – 99**	**100 – 149**	**150 – 199**	**200 – 249**	**250 – 299**	**300 – 349**
Frequency	6	12	18	9	4	1

Compute the mean, median, mode, quartile deviation and standard deviation. [Answers: $\bar{x} \approx 170.5$; Median ≈ 166.9; Mode ≈ 163.6; QD ≈ 28.7; $s \approx 58.1$]

10. The prices (in thousand rupees) of a particular brand of mobile phone at eight retail outlets in Rawalpindi are 32, 35, 33, 36, 34, 38, 37, 35. Compute the range, the sample variance, the sample standard deviation and the coefficient of variation. [Answers: $R = 6$; $s^2 \approx 3.71$; $s \approx 1.93$; CV $\approx$ 5.48%]

11. The annual sales growth of a textile company over six consecutive years was 4%, 6%, 9%, 3%, 7% and 5%. Find the average annual growth rate using the geometric mean. [Answer: GM $\approx$ 5.65% per year]

12. A student of BS Education at AIOU obtained the following marks in five courses, with the indicated credit hours:

Course	**Credit Hours**	**Marks Obtained**
Statistics	4	78
Economics	3	82
Sociology	3	69
English	2	75
Islamic Studies	2	88

Compute the weighted mean marks for the semester. [Answer: $\bar{x}_w \approx 77.57$]

13. The ages (in years) of nine teachers in a private school in Islamabad are 28, 35, 42, 31, 55, 38, 46, 29, 40. Find Q_1, Q_3, the quartile deviation and the coefficient of quartile deviation. [Answers: $Q_1 = 30$; $Q_3 = 44$; QD = 7; Coefficient of QD ≈ 0.189]

14. The hourly output of a machine in a manufacturing unit in Faisalabad was recorded for twelve consecutive hours as 45, 48, 52, 50, 47, 46, 53, 49, 51, 48, 50, 47 units. Compute the mean, the standard deviation, the mean deviation about the mean, and the coefficient of variation. [Answers: $\bar{x} \approx 48.83$; $s \approx 2.37$; MD(mean) ≈ 1.93; CV $\approx 4.86\%$]

15. Two groups of students were given an aptitude test. Group A had 40 students with a mean of 65 and a standard deviation of 8, while Group B had 50 students with a mean of 70 and a standard deviation of 12. Which group is more consistent? Also compute the combined mean of the two groups. [Answers: Group A is more consistent (CV 12.31% vs 17.14%); combined mean ≈ 67.78]

16. The following data show the number of customers arriving per hour at a branch of Habib Bank in Islamabad on a certain working day: 35, 42, 38, 45, 40, 47, 39, 44, 36, 41, 43, 46, 37, 42, 40. Compute the mean, the median, the sample variance, the sample standard deviation and Karl Pearson's coefficient of skewness. [Answers: $\bar{x} \approx 41.0$; Median = 41; $s^2 \approx 12.43$; $s \approx 3.53$; $Sk_P \approx 0$]

17. For a frequency distribution we are given mean = 45, median = 47 and standard deviation = 8. Compute Karl Pearson's second coefficient of skewness and comment on the shape of the distribution. [Answer: $Sk_P = -0.75$; moderately negatively skewed]

UNIT 2: SAMPLING AND SAMPLING DISTRIBUTIONS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Distinguish clearly between a population and a sample, and between a parameter and a statistic, and explain why sampling is essential in nearly every real-world research project.
2. Describe and apply the principal probability sampling methods — simple random, stratified, systematic, cluster and multi-stage sampling — and explain how each differs from the non-probability methods used in opinion polls and journalistic surveys.
3. Explain what is meant by sampling and non-sampling errors, and describe the practical steps a researcher can take to keep both kinds of error within acceptable limits.
4. Define a sampling distribution, derive the mean and standard error of the sampling distribution of the sample mean and the sample proportion, and state the Central Limit Theorem in plain language.
5. Use the sampling distribution to make basic inferences — point and interval estimates — about a population mean or proportion, and explain why doubling the sample size does not halve the standard error.

2.1 INTRODUCTION

In Unit 1 we developed the descriptive tools that allow us to summarize a data set in a compact way. We learned how to compute measures of central tendency, measures of dispersion, quantiles and measures of shape. All of these techniques, however, were applied to whatever data happened to be in front of us. We never paused to ask the most important practical question of all: where did the data come from in the first place, and how should they have been collected? That question is the subject of the present unit, and the answer to it lies at the heart of all modern statistical inference.

In nearly every real-world social-science investigation, the researcher is interested in some large group of people, households, firms, schools, hospitals or events — say, all adult voters in Pakistan, all primary-school children in Punjab, all small businesses in Karachi, or all patients admitted to a particular hospital during a given year. This large group is called the population. Sometimes the population is small enough that the researcher can examine every single member of it; this is called a complete enumeration or a census. Most of the time, however, the population is far too large for a complete enumeration to be practical. The cost would be enormous, the time involved would be very long, the workload on the field staff would be unmanageable, and in many cases the population is so vast or so widely scattered that reaching every member is simply impossible. Even when a complete enumeration is theoretically possible, it is rarely the most efficient way to obtain the desired information.

The alternative to a complete enumeration is to examine only a carefully chosen part of the population — called a sample — and then to use the information contained in the sample to draw conclusions, or inferences, about the population as a whole. The branch of statistics that draws such conclusions is called inferential statistics, and it is the central concern of this entire course. Sampling, however, is not as simple as just picking a few people at random from the street. To produce inferences that are reliable and free of bias, the sample must be drawn in a carefully controlled way, and the researcher must understand exactly how the act of sampling introduces uncertainty into the conclusions. The mathematical tool that captures this uncertainty is called the sampling distribution, and it is the single most important idea in inferential statistics.

The present unit introduces all the basic ideas about sampling and sampling distributions that the student will need for the remaining units of the course. We begin with the difference between a population and a sample, and between a parameter and a statistic. We then describe the main sampling methods used in practice and the kinds of error that sampling can produce. After that we develop the concept of a sampling distribution and state the Central Limit Theorem, which is the theoretical foundation for almost all the inferential techniques that follow. Finally we look briefly at point and interval estimation, and we examine the relationship between sample size and the precision of an estimate.

The student is strongly advised to work through every solved example carefully. The ideas in this unit are abstract — more abstract, certainly, than the descriptive measures of Unit 1 — but they are also extremely important. A student who understands sampling and sampling distributions thoroughly will find the rest of the course much easier; a student who skims this unit will struggle with everything that follows.

2.2 POPULATION AND SAMPLES

We begin with two of the most fundamental terms in statistics: the population and the sample. These two words have a precise technical meaning that is somewhat narrower than their everyday usage, and it is essential that the student commit the technical definitions to memory.

2.2.1 Definition of a Population

A population, in the statistical sense, is the entire collection of individuals, objects or measurements about which information is desired. The members of the population are called the units (or sometimes the elementary units, or simply the elements) of the population. The population must be defined precisely by the researcher at the start of the investigation, because vague definitions lead to vague conclusions. For example, a researcher who wants to study "students in Pakistan" must specify exactly what is meant by this phrase: only those enrolled in formal degree programs, or also those in informal education? Only full-time students, or also part-time? Only those in the public sector, or also in private institutions? Only those above a certain age, or all? Until the population is defined precisely, no meaningful sample can be drawn from it.

Populations are commonly classified along two dimensions. The first is whether the population is finite or infinite. A finite population contains a definite, countable number of units — for example, all 220 students enrolled in a particular degree program at Allama Iqbal Open University in a given semester. An infinite population, on the other hand, contains an unlimited number of units. The number of possible outcomes when a coin is tossed indefinitely many times, the number of measurements that could conceivably be made of the boiling point of water under identical conditions, and the number of customers who could potentially walk into a particular shop over an unlimited period of time are all examples of (theoretically) infinite populations. In practice, "infinite" usually means simply that the population is so large that, for the purposes of sampling, it might as well be infinite.

The second classification is between an existent population and a hypothetical population. An existent population is one whose units actually exist at the time of the study — for example, all currently registered voters in Lahore. A hypothetical population is one whose units are conceptual or future-oriented — for example, all the children who will ever be born in a particular town, or all the cars that will ever be produced by a certain factory under its current production process. Hypothetical populations are common in industrial quality control and in scientific experiments, where each measurement is thought of as a draw from an indefinitely large pool of possible measurements.

2.2.2 Definition of a Sample

A sample is any subset of a population. In nearly every practical investigation the sample is much smaller than the population from which it is drawn. The number of units in the sample is called the sample size and is universally denoted by n; the number of units in the population is called the population size and is denoted by N. For finite populations we always have $n \leq N$, and in most real-world studies n is very much smaller than N — sometimes only a fraction of one per cent of N. The art of sampling is to choose this small subset in such a way that it represents the population fairly, so that conclusions drawn from the sample can safely be generalized to the whole population.

A sample drawn in such a way that every unit of the population has a known, non-zero probability of being included is called a probability sample, and the technique used to draw it is called a probability sampling method or random sampling method. A sample drawn in any other way is called a non-probability sample. Probability samples are the gold standard in statistical research because they are the only kind of samples for which the formal mathematical tools of statistical inference can be applied with full validity. Non-probability samples are nevertheless widely used in journalism, marketing, exploratory research and qualitative studies, where the goals are different and the formal probability framework is less important.

2.2.3 The Population–Sample Relationship

The relationship between a population and a sample is best visualized as a circle (the population) containing a smaller dot (the sample). The researcher cannot afford to examine every member of the population, so he or she draws a small subset, makes the desired measurements, and then uses those sample measurements to make a statistical guess about the corresponding quantity in the population.

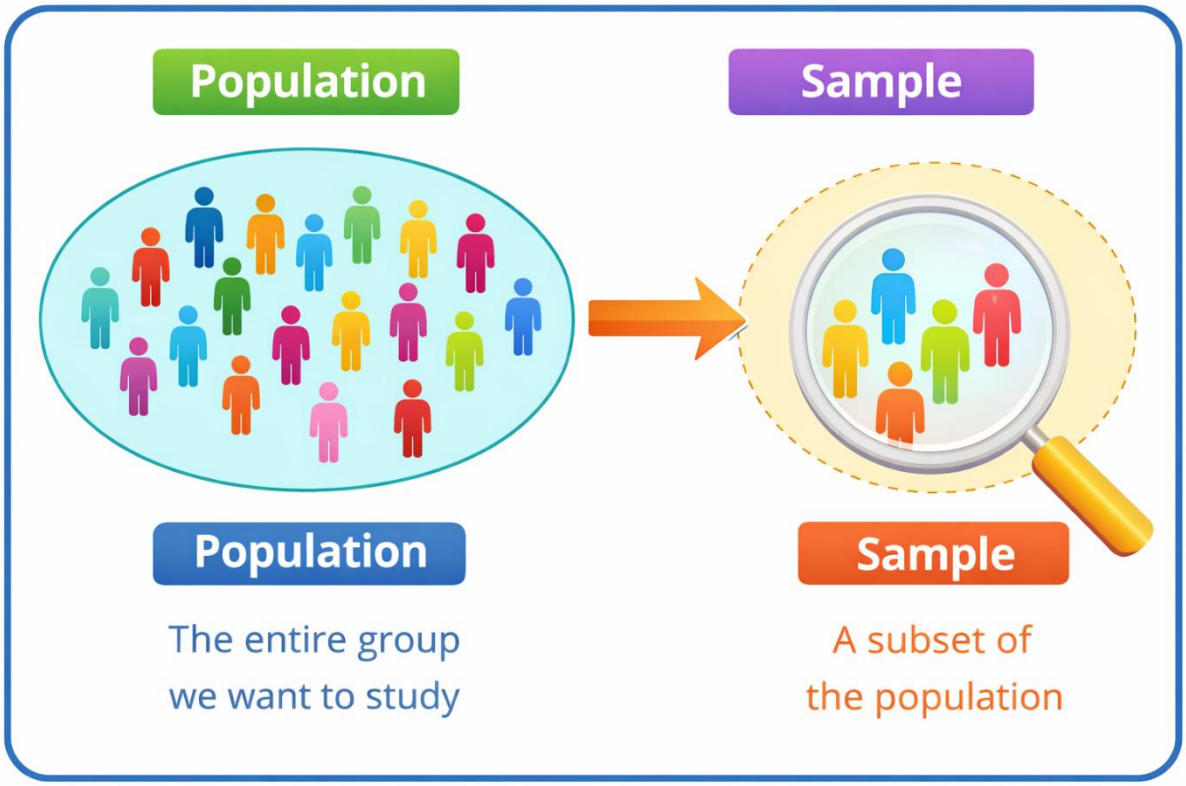


Figure 2.1 Visual representation of Population and Sample

The whole logic of statistical inference rests on the idea that, if the sample has been drawn properly, the things we observe in the sample are likely to be similar to the things that are true of the population — although not exactly the same, because of what we shall later call sampling error.

2.3 PARAMETERS AND ESTIMATES

Closely tied to the distinction between a population and a sample is the equally important distinction between a parameter and a statistic. We met both terms briefly in Unit 1; in this section we develop them in more detail.

2.3.1 Parameters

A parameter is any numerical characteristic of a population. The mean, the variance, the standard deviation, the proportion of units having a particular property, the correlation between two variables, and the median are all examples of parameters when they are computed for the whole population. By universal convention, parameters are denoted by lower-case Greek letters, the most common being:

Parameter	Symbol	Definition
Population mean	μ	Arithmetic mean of all N population values
Population variance	σ^2	Average squared deviation from μ in the population
Population standard deviation	σ	Square root of σ^2

Parameter	Symbol	Definition
Population proportion	π or P	Fraction of population units with a given attribute
Population correlation	ρ	Correlation between two variables in the population

The crucial point about parameters is that they are fixed but unknown constants. They are fixed because they are characteristics of the population, which itself is fixed at any given moment in time. They are unknown because the population is too large to be observed completely; if it could be observed completely, there would be no need for inferential statistics at all. The fundamental problem of statistical inference is to make educated guesses about these unknown population parameters on the basis of information contained in a sample.

2.3.2 Statistics and Estimates

A statistic is any numerical quantity computed from a sample. The sample mean $\bar{x}$, the sample variance s^2, the sample standard deviation s, the sample proportion $\hat{p}$, and the sample correlation coefficient r are all examples of statistics. By convention, statistics are denoted by lower-case Latin letters, in deliberate contrast to the Greek letters used for parameters. The contrast is summarized once more in the following table:

Quantity	Population (Greek)	Sample (Latin)
Number of observations	N	N
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	Π	$\hat{p}$
Correlation	ρ	r

Unlike parameters, statistics are not fixed: their values change every time we draw a new sample. If we draw two different samples of the same size from the same population, the two samples will, in general, give two different values of the sample mean, of the sample standard deviation, and so on. This variability is unavoidable, and it is precisely what makes statistical inference a non-trivial subject. We will see in Section 2.12 that this sample-to-sample variability is captured by the sampling distribution of the statistic in question.

When a statistic is used as a guess for the value of an unknown parameter, it is called an estimator of that parameter, and the particular numerical value it takes for a given sample is called an estimate. The sample mean $\bar{x}$ is the standard estimator of the population mean μ; the sample variance s^2 is the standard estimator of the population variance σ^2; the sample proportion $\hat{p}$ is the standard estimator of the population proportion π. The student should pause here and re-read this paragraph carefully: estimator, estimate and parameter are three different things, and confusing them is a common source of error in introductory courses.

Example 2.1 (Identifying Populations, Samples, Parameters and Statistics)

A researcher at Allama Iqbal Open University wants to study the average monthly household income of families living in District Rawalpindi. There are about 850,000 households in the district. The researcher selects 600 households at random and finds that the mean monthly income in these

600 households is Rs. 64,250 with a standard deviation of Rs. 18,400. Identify the population, the sample, the parameter of interest, and the corresponding sample statistics.

Solution.

1. The population is the set of all 850,000 households in District Rawalpindi (so $N \approx 850{,}000$).
2. The sample is the set of 600 households actually examined (so $n = 600$).
3. The parameter of interest is the mean monthly household income μ in the entire district. This is a fixed but unknown number.
4. The sample statistics are $\bar{x} = 64{,}250$ rupees and $s = 18{,}400$ rupees. These are the numerical estimates that the researcher will use to make inferences about μ.

Notice the deliberate use of Greek and Latin letters: μ for the unknown population mean, $\bar{x}$ for the known sample mean. This habit, once acquired, will save the student from many errors in later units.

2.4 INTRODUCTION TO STATISTICAL INFERENCE

Statistical inference is the process of drawing conclusions about the population on the basis of information contained in a sample. The conclusions are necessarily uncertain — because they are based only on a small part of the population — and one of the most important tasks of statistics is to express that uncertainty in a precise mathematical way. There are two broad branches of statistical inference, namely estimation and testing of hypotheses, and we will study both branches in some detail in Units 3 to 6.

2.4.1 Estimation

In estimation, the goal is to make an educated guess about the value of an unknown population parameter. There are two distinct kinds of estimation. In point estimation, the researcher gives a single numerical value as the best guess for the parameter. For example, on the basis of a sample of 600 households, we might estimate that the mean monthly household income in District Rawalpindi is Rs. 64,250. This single number is called a point estimate. In interval estimation, the researcher gives an entire range of plausible values for the parameter, together with a statement of how confident he or she is that the true value lies within that range. For example, we might estimate that the mean monthly household income is between Rs. 62,780 and Rs. 65,720, and that we are 95% confident in this statement. Such an interval is called a confidence interval, and the percentage attached to it is called the confidence level. We will study both kinds of estimation in detail in Unit 3.

2.4.2 Testing of Hypotheses

In hypothesis testing, the researcher does not start with the goal of guessing a parameter value but with a specific claim about the parameter that he or she wants to verify or refute. For example, a manufacturer of energy-saving bulbs may claim that the average life of his bulbs is 1,500 hours. A consumer-protection agency, skeptical of this claim, draws a sample of 50 bulbs, measures their lifetimes, computes the sample mean, and then asks: "given the sample mean we observed, is the manufacturer's claim of 1,500 hours believable, or is it likely that the true mean is something different?" The formal procedure used to answer such a question is called a test of hypothesis, and the various forms it takes (one-sample tests, two-sample tests, tests of proportions, tests of variances, tests of independence, and so on) will occupy us in Units 4 to 6.

The two branches of inference, estimation and hypothesis testing, are closely related and are based on the same underlying theory. In both cases, the bridge between the sample and the population is the sampling distribution of the statistic being used. We turn to that idea in Section 2.12. Before we do, we need to discuss why sampling is so important and how to do it properly.

2.5 IMPORTANCE OF SAMPLING IN STATISTICS

Why do we sample at all? Why not just examine the entire population, which would obviously give us the most accurate possible information? The answer is that, even when a complete enumeration is theoretically possible, sampling has so many practical advantages that it is almost always the preferred method. The main reasons for using a sample rather than a complete enumeration may be summarized under five headings.

2.5.1 Reduced Cost

A complete enumeration of a large population is enormously expensive. Consider the cost of conducting a national census in a country the size of Pakistan, with a population of more than 240 million people: hundreds of thousands of enumerators have to be hired and trained, millions of forms have to be printed, transport has to be arranged to every village and every neighborhood, the data have to be entered into computers, checked, cleaned, processed and analyzed. The total cost runs into many billions of rupees. A well-designed sample survey covering, say, fifty thousand households can produce nearly the same information at a tiny fraction of the cost. For most research purposes the small loss in precision that comes from sampling is more than compensated by the enormous saving in money.

2.5.2 Greater Speed

A sample can be drawn, processed and analyzed in a fraction of the time required for a complete enumeration. In opinion polls during an election campaign, in market research before launching a new product, in epidemiological studies during the early stages of a disease outbreak — in all these situations, results that arrive too late are useless. A sample of two thousand voters can be polled and the results published in a single day; a complete enumeration of all voters would take months at best. Speed is often the single most important advantage of sampling over complete enumeration.

2.5.3 Greater Accuracy

Surprising as it may sound, a sample is often more accurate than a complete enumeration. The reason is that, in a sample survey, the field staff can be smaller, better trained, more closely supervised and more carefully monitored. A small team of well-trained interviewers, each handling a few hundred questionnaires, will commit far fewer errors than the army of poorly trained enumerators that a complete enumeration would require. The errors of measurement, recording, classification and computation in a complete enumeration can easily exceed the so-called sampling error of a well-designed sample, with the result that the sample actually gives the more accurate answer.

2.5.4 Greater Scope

A sample makes it possible to ask more questions and to ask them in greater depth than is possible in a complete enumeration, where time pressure usually forces the questionnaire to be short and superficial. In a sample survey of two thousand households, an interviewer can comfortably spend an hour or more in each household, asking detailed questions about income, expenditure, education, health, employment, family

relationships and aspirations. The same depth of questioning would be impossible in a complete enumeration of millions of households.

2.5.5 Necessity

Sometimes a complete enumeration is not merely inconvenient but actually impossible. If the population is infinite, by definition we cannot examine every member of it. If the act of measuring a unit destroys it — for example, when testing the breaking strength of fibres, the lifetime of light bulbs or the explosive power of fireworks — we obviously cannot test every unit in the population, because there would then be nothing left to use. In all such situations sampling is not an option but a necessity.

2.5.6 When is a Complete Enumeration Preferred?

In spite of all these advantages of sampling, there are still a few situations in which a complete enumeration is preferred. The first is when the population is small enough that the cost of a complete enumeration is no greater than the cost of a sample survey — for example, when studying the staff of a single school or the patients in a single ward of a hospital. In such cases there is little point in sampling, because every unit can be examined directly. The second situation is when extremely high accuracy is required for legal, administrative or constitutional reasons — for example, in a national population census, which is usually mandated by law and forms the basis of constituency boundaries, parliamentary seats and inter-provincial revenue distribution. In such cases the small additional precision provided by a complete enumeration is worth the great additional cost. The third situation is when information is required not only for the country as a whole but for every small administrative unit (every district, tehsil and union council) — a sample survey large enough to give precise estimates at this level of disaggregation would be almost as expensive as a full census. In all other situations — and that is the vast majority of social-science research — sampling is the right choice.

2.6 RANDOM (PROBABILITY) SAMPLING METHODS

A sampling method is called random or probabilistic if every unit of the population has a known, non-zero probability of being included in the sample, and if that probability is determined by some chance mechanism rather than by the personal preferences of the researcher. Random sampling is the foundation of inferential statistics for two closely related reasons. First, it eliminates the personal bias of the researcher: the units to be included in the sample are chosen by a coin, a table of random numbers, or a computer-generated random sequence, not by the researcher's convenience or by his preconceived ideas about which units are interesting. Second, only random samples allow us to apply the formal mathematical theory of probability to the sampling distribution, and thus to attach precise probabilities and confidence levels to our inferences about the population.

There are several different methods of random sampling. The most important ones are simple random sampling, stratified random sampling, systematic sampling, cluster sampling and multi-stage sampling. Each has its own strengths and weaknesses, and the choice between them depends on the nature of the population, on the practical constraints under which the survey is being conducted, and on the kind of information that the researcher wishes to obtain. We discuss each method in turn in the next four sections.

2.7 SIMPLE RANDOM SAMPLING

Simple random sampling is the simplest and most fundamental of all the random sampling methods. A simple random sample of size n drawn from a population of size N is a sample selected in such a way that every possible sample of size n has exactly the same probability of being drawn. An equivalent way of stating the

same thing is that, at each draw, every unit of the population that has not yet been selected has the same probability of being selected next. Simple random sampling is the building block from which the more sophisticated sampling methods are constructed.

2.7.1 Sampling With and Without Replacement

Simple random sampling can be carried out in two slightly different ways. In sampling with replacement, each unit selected is "returned" to the population before the next unit is drawn, so that the same unit can in principle be selected more than once. In sampling without replacement, each unit selected is set aside and cannot be drawn again. Sampling without replacement is by far the more common in practice, because it is wasteful to interview the same household twice, but sampling with replacement has slightly simpler mathematical properties and is often used in theoretical derivations.

The total number of possible samples is different in the two cases. For sampling with replacement, the number of possible samples of size n drawn from a population of size N is:

$$Number\ of\ samples\ (with\ replacement)\ =\ N^n \qquad (2.1)$$

For sampling without replacement, the number of possible samples is given by the binomial coefficient:

$$Number\ of\ samples\ (without\ replacement)\ = \binom{N}{n} = \frac{N!}{n!\,(N-n)!} \qquad (2.2)$$

where N! denotes the factorial of N, that is, N! = N × (N − 1) × (N − 2) × ⋯ × 2 × 1, with the convention 0! = 1.

Example 2.2 (Counting Possible Samples)

A teacher at a school in Islamabad has a class of N = 5 students. He wants to draw a simple random sample of n = 2 students for an oral interview. (a) How many distinct samples of size 2 can be drawn without replacement? (b) How many can be drawn with replacement?

Solution.

(a) Without replacement, using equation (2.2):

$$\binom{5}{2} = \frac{5!}{2!\,(5\ -\ 2)!} = \frac{120}{2\ \times\ 6}\ =\ 10.$$

There are 10 distinct possible samples of size 2 that can be drawn without replacement from this class of 5 students.

(b) With replacement, using equation (2.1):

$$5^2\ =\ 25.$$

There are 25 possible samples of size 2 that can be drawn with replacement, because each of the two draws independently has 5 possible outcomes.

Example 2.2(b) (Listing All Possible Samples and Their Means)

Suppose the five students of Example 2.2 obtained the following scores in a quiz: A = 6, B = 8, C = 10, D = 12 and E = 14. List all 10 possible samples of size n = 2 drawn without replacement, and compute

the sample mean for each. Verify that the mean of these 10 sample means is exactly equal to the population mean μ.

Solution. The population mean is $\mu = (6 + 8 + 10 + 12 + 14) / 5 = 50 / 5 = 10$. The 10 possible samples and their means are:

Sample	Values	Sample Mean $\bar{x}$
(A, B)	(6, 8)	7
(A, C)	(6, 10)	8
(A, D)	(6, 12)	9
(A, E)	(6, 14)	10
(B, C)	(8, 10)	9
(B, D)	(8, 12)	10
(B, E)	(8, 14)	11
(C, D)	(10, 12)	11
(C, E)	(10, 14)	12
(D, E)	(12, 14)	13
Total	—	100

The mean of the 10 sample means is 100 / 10 = 10, which equals the population mean $\mu = 10$ exactly. This is a numerical verification of the general result $\mu_{\bar{x}} = \mu$ that we established in equation (2.4): on average, the sample mean equals the population mean. Notice also that the individual sample means range from 7 to 13 — that is, they vary by ± 3 around the true value of 10. This sample-to-sample variability is precisely what the standard error $\sigma_{\bar{x}}$ measures.

2.7.2 How to Draw a Simple Random Sample in Practice

In practice, a simple random sample is drawn using a table of random numbers, a lottery method, or — most commonly today — a computer. The procedure consists of the following steps:

1. Prepare a complete list of all units in the population, called the sampling frame, and assign a unique identification number to each unit, from 1 to N.
2. Decide on the sample size n, taking into account the precision required, the budget available, and the variability of the characteristic being studied.
3. Use a table of random numbers, a lottery, or a computer program (such as the RAND function in Excel, or the runif() function in R) to generate n distinct random integers between 1 and N.
4. Select the units whose identification numbers correspond to those random integers. The resulting set of n units is the simple random sample.

2.7.3 Advantages and Limitations

Advantages. Simple random sampling is easy to understand and easy to implement when the sampling frame is small. It produces an unbiased sample, and the formulas for sampling distributions, standard errors

and confidence intervals take their simplest form for simple random samples. It is therefore the natural starting point for any first course on statistical inference.

Limitations. Simple random sampling has two important limitations. First, it requires a complete and accurate list of the entire population, which may be expensive or impossible to obtain — imagine trying to compile a list of every adult voter in Pakistan before drawing a sample. Second, when the population is geographically scattered, a simple random sample may include units that are very far apart, making the field work very expensive. The sampling methods of the next sections are designed to overcome one or both of these limitations.

2.8 STRATIFIED AND PROPORTIONAL STRATIFIED SAMPLING

Suppose the population we wish to study is not internally homogeneous but consists of several clearly distinct sub-groups. For example, the population of all employees of a large bank may consist of three quite different sub-groups: senior managers, middle managers and clerical staff. The mean salary, the educational background, the age structure and the working hours of these three sub-groups may be very different from one another. A simple random sample drawn from the bank as a whole might, by bad luck, contain mostly clerical staff and very few managers, leading to a biased picture of the whole organization. Stratified random sampling is designed to avoid such an outcome.

2.8.1 The Idea of Stratification

In stratified random sampling, the population is first divided into a number of non-overlapping sub-groups called strata, in such a way that each unit of the population belongs to exactly one stratum. The strata are constructed so that, within any single stratum, the units are as similar to one another as possible (homogeneous within), while the strata themselves are as different from one another as possible (heterogeneous between). A simple random sample is then drawn independently from every stratum, and these stratum-level samples are combined to form the overall stratified sample.

Common bases for stratification in social-science research include gender (male, female), age group (young, middle-aged, elderly), education level (illiterate, primary, secondary, tertiary), region (urban, rural), province (Punjab, Sindh, KPK, Balochistan), occupation (agricultural, industrial, service) and income class (low, middle, high). The choice of strata depends on the variable being studied: the strata should be defined in such a way that they are likely to differ from one another with respect to that variable.

2.8.2 Allocating the Sample to the Strata

Once the strata have been defined, the researcher must decide how many units to draw from each stratum. Two main allocation schemes are in common use.

Equal allocation. Under equal allocation, the same number of units is drawn from every stratum, regardless of how big or small the stratum is. If there are k strata and a total sample size of n, then n/k units are drawn from each stratum. Equal allocation is simple to implement, but it can give too much weight to small strata and too little to large ones.

Proportional allocation. Under proportional allocation, the number of units drawn from each stratum is proportional to the size of that stratum. If the i-th stratum contains N_i units out of the total population of N units, then the number of units drawn from the i-th stratum, denoted n_i, is given by:

$$n_i = n \cdot \frac{N_i}{N} \quad , \quad i = 1, 2, \ldots, k. \qquad (2.3)$$

Proportional allocation gives every unit of the population an equal chance of being included in the final sample, just as in simple random sampling, but with the added guarantee that every stratum is properly represented in proportion to its size in the population. For this reason proportional stratified sampling is the most popular form of stratification in social-science research.

Example 2.3 (Proportional Stratified Sampling)

A medium-sized college in Lahore has a total of N = 2,000 students enrolled in three faculties: Faculty of Arts (1,000 students), Faculty of Sciences (700 students), and Faculty of Commerce (300 students). The Vice-Principal wants to draw a stratified random sample of n = 200 students using proportional allocation, with the three faculties as strata. How many students should be drawn from each faculty?

Solution. Applying equation (2.3) to each of the three strata:

$$n_{Arts} = 200 \cdot \frac{1000}{2000} = 100\ students,$$

$$n_{Sciences} = 200 \cdot \frac{700}{2000} = 70\ students,$$

$$n_{Commerce} = 200 \cdot \frac{300}{2000} = 30\ students.$$

The total is 100 + 70 + 30 = 200, as required. The Vice-Principal will then draw a simple random sample of 100 students from the Faculty of Arts, an independent simple random sample of 70 students from the Faculty of Sciences, and an independent simple random sample of 30 students from the Faculty of Commerce. The combined sample of 200 students is the proportional stratified random sample.

2.8.3 Advantages of Stratified Sampling

Stratified random sampling has three important advantages over simple random sampling. First, it guarantees that every stratum is represented in the final sample, which prevents the kind of accidental imbalance that simple random sampling sometimes produces. Second, when the strata are constructed sensibly (homogeneous within and heterogeneous between), the stratified estimate of a population parameter has a smaller standard error than the corresponding simple-random-sample estimate of the same size, so stratification yields more precise inferences. Third, stratified sampling allows the researcher to draw separate inferences for each stratum, which can itself be valuable — a study of student satisfaction, for example, may want to compare the three faculties as well as to estimate the overall satisfaction level for the college as a whole.

Example 2.3(c) (Estimating a Population Mean from a Stratified Sample)

Continuing Example 2.3, suppose that the proportional stratified sample of 200 students gave the following sample means for the average number of hours spent on study per week:

Stratum	Stratum size N_i	Sample size n_i	Sample mean $\bar{x}_i$ (hours)
Arts	1000	100	14.5

Sciences	700	70	18.2
Commerce	300	30	16.0
Total	2000	200	—

Estimate the overall mean number of study hours per week for the entire college population.

Solution. The estimator of the population mean from a stratified sample is the weighted average of the stratum sample means, with weights proportional to the stratum sizes:

$$\bar{x}_{st} = \frac{\sum_{i=1}^{k} N_i \bar{x}_i}{N}$$

$$\bar{x}_{st} = \frac{1000(14.5) + 700(18.2) + 300(16.0)}{2000} = \frac{14500 + 12740 + 4800}{2000} = \frac{32040}{2000}$$
$$= 16.02\ hours.$$

The estimated overall mean number of study hours per week is approximately 16 hours per student. Notice that, because of proportional allocation, this stratified estimate works out to exactly the same value as we would have obtained by computing the unweighted mean of all 200 sample observations directly. The advantage of stratification is not so much in the value of the estimate itself, but in the increased precision (smaller standard error) of that estimate compared with a simple random sample of the same total size.

2.9 OTHER SAMPLING PROCEDURES

2.9.1 Systematic Sampling

In systematic sampling, the researcher first prepares a complete list of the population, then chooses a starting point at random within the first few units of the list, and finally selects every k-th unit thereafter. The number k is called the sampling interval and is computed as k = N/n (rounded to the nearest integer). For example, if a population of N = 1,000 voters is to be sampled with n = 50, the sampling interval is k = 1,000 / 50 = 20. The researcher chooses a random integer between 1 and 20 (say 7) and then selects voters number 7, 27, 47, 67, 87, …, 987 — a total of 50 voters.

Systematic sampling is much faster and easier to implement than simple random sampling, especially when the population is already organized into a physical or alphabetical list (such as the customers of a bank, the patients of a hospital, or the houses on a street). It is, however, vulnerable to a particular kind of bias: if the population list itself contains a hidden cycle of length k or a multiple of k, the systematic sample will pick up that cycle and produce a misleading result. For example, if the houses on a street are numbered 1, 2, 3, … and every tenth house is a corner house at an intersection, then a systematic sample with interval k = 10 will consist entirely of corner houses, which are likely to differ systematically from the others. Provided no such hidden cycle exists, systematic sampling gives results that are very nearly as good as simple random sampling but at much lower cost.

Example 2.3(b) (Systematic Sampling)

A market researcher in Lahore wishes to survey n = 60 shoppers from a department store that admits about N = 1,800 shoppers per day. Describe how she could use systematic sampling, and identify the first five shoppers she would select.

Solution. The sampling interval is k = N/n = 1,800/60 = 30. The researcher first chooses a random integer between 1 and 30 — say, by drawing a slip of paper from a bag — and suppose she gets 14. She then interviews shopper number 14 (counting in the order they enter the store), then shopper number 14 + 30 = 44, then 44 + 30 = 74, then 74 + 30 = 104, then 104 + 30 = 134, and so on, until she has interviewed 60 shoppers in total. The first five shoppers in her sample are therefore numbers 14, 44, 74, 104 and 134. Notice that, because the random starting point was 14, every shopper in the population had exactly a 1/30 chance of being included in the sample, which makes this a genuine probability sample.

2.9.2 Cluster Sampling

Cluster sampling is the natural choice when the population is geographically scattered and a simple random sample would be prohibitively expensive to interview. The population is divided into a large number of small groups called clusters, where each cluster is geographically compact and contains many units. A small number of clusters is then selected at random, and within each selected cluster, every unit (or sometimes a sub-sample of units) is examined. The clusters in cluster sampling are typically defined by geography: in a national household survey, for instance, each cluster might be a single village in rural areas or a single city block in urban areas.

The crucial difference between stratified sampling and cluster sampling lies in what is done with the strata or clusters once they have been defined. In stratified sampling, every stratum is sampled, but only a few units are drawn from each. In cluster sampling, only a few clusters are sampled, but within each chosen cluster all (or most) units are examined. Stratified sampling aims to maximize the precision of the estimate by ensuring that every sub-group is represented; cluster sampling aims to minimize the cost by concentrating field work in a small number of geographic locations. The two methods are therefore complementary, not competing, and they are often combined in large household surveys.

2.9.3 Multi-Stage Sampling

Most large household surveys conducted in Pakistan and elsewhere use a combination of the methods we have just described, in what is called multi-stage sampling. In a typical two-stage design, the country is first divided into a large number of primary sampling units (such as villages or city blocks), and a sample of these primary units is drawn. Within each chosen primary unit, a list of all the households is then prepared, and a simple random sample (or sometimes a systematic sample) of households is drawn from each list. In a three-stage design, the chosen primary units are subdivided into smaller secondary units (such as neighborhoods), a sample of secondary units is drawn within each chosen primary unit, and households are then sampled within the chosen secondary units. The Pakistan Bureau of Statistics uses such a multi-stage design in its periodic Household Integrated Economic Survey (HIES).

Multi-stage designs are extremely flexible and combine many of the advantages of stratified and cluster sampling, but the formulas for the standard error of the resulting estimates are correspondingly more complicated. We will not go into those formulas in this introductory course; the interested student is referred to Cochran (1977) for a complete treatment.

2.10 NON-PROBABILITY SAMPLING METHODS

All the sampling methods we have described so far are random or probability sampling methods. There is, however, a second broad family of sampling methods in which the units to be included in the sample are chosen not by chance but by the personal judgement of the researcher. These are called non-probability sampling methods, and they are widely used in journalism, marketing research, opinion polling, exploratory studies and qualitative research. Non-probability samples are easier and cheaper to collect than random samples, but they suffer from a fundamental weakness: there is no formal mathematical theory for computing the standard error or the confidence level of estimates based on them, and there is no guarantee that the sample is representative of the population. Inferences drawn from non-probability samples should therefore always be regarded with caution.

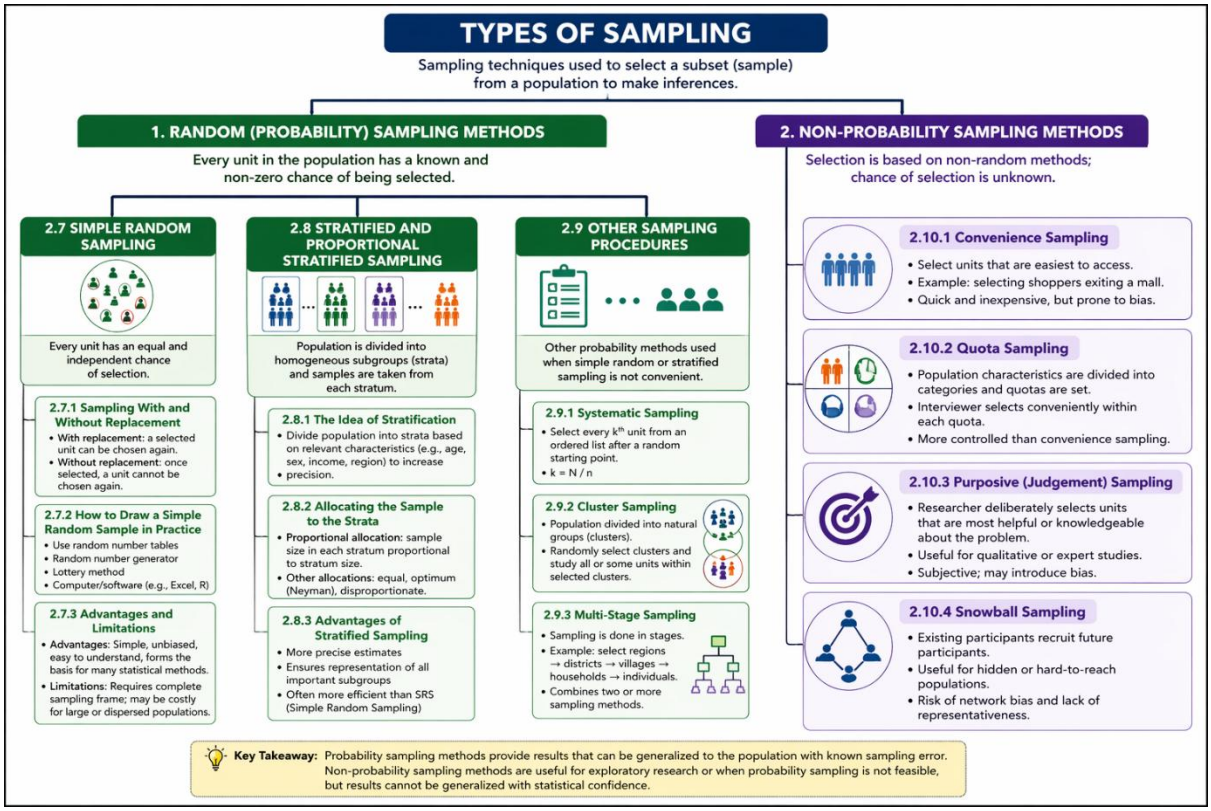


Figure 2.2: A Comparative Overview of Probability and Non-Probability Methods

2.10.1 Convenience Sampling

Convenience sampling — sometimes called accidental sampling — consists of choosing whichever units happen to be most easily available. The classic example is the man-in-the-street interview, in which a journalist interviews the first ten or twenty passers-by who happen to be in the right place at the right time. Convenience samples are quick and cheap, but they are notoriously biased: the people who happen to be available at any particular time and place are usually quite different from the population as a whole.

2.10.2 Quota Sampling

Quota sampling is the non-probability counterpart of stratified sampling. The researcher first divides the population into a number of categories (quotas) based on characteristics such as age, sex, occupation and income, and then instructs the field staff to interview a fixed number of respondents in each category. Within each category, however, the actual respondents are chosen by the convenience or judgement of the interviewer rather than by chance. Quota sampling is widely used in commercial market research because it

produces samples that look balanced with respect to a few key demographic variables, but the underlying selection within each quota is still a convenience sample, and the same warnings apply.

2.10.3 Purposive (Judgement) Sampling

In purposive or judgement sampling, the researcher uses his or her own expert knowledge to deliberately choose units that are believed to be especially informative, typical or interesting. For example, a sociologist studying poverty in Pakistan may deliberately select a few villages thought to be representative of different geographic and economic conditions. Purposive sampling is useful in qualitative research and in pilot studies, but it gives no objective measure of how representative the chosen units actually are.

2.10.4 Snowball Sampling

Snowball sampling is used when the population of interest is difficult to identify or to reach directly — for example, in studies of people belonging to small or hidden communities, or to specialized professions. The researcher begins with a few initial respondents who meet the criterion of interest, and then asks each of them to identify other people they know who also meet the criterion. The sample grows like a snowball rolling downhill. Snowball sampling is the only feasible method in some kinds of research, but the resulting sample is heavily concentrated within a few interconnected social networks and is rarely representative of the broader population.

2.11 ERRORS IN SAMPLING

Whenever we use a sample instead of a complete enumeration, our estimate of the population parameter is bound to differ from the true value of that parameter, simply because the sample contains less information than the entire population. The total error in any survey is conventionally divided into two broad categories: sampling errors and non-sampling errors. The student should be familiar with both kinds, because each calls for a different remedy.

2.11.1 Sampling Errors

Sampling error is the error that arises purely from the fact that we have observed only a sample, not the whole population. Even if the sample is drawn perfectly at random, and even if every measurement is recorded perfectly, the sample mean will almost never be exactly equal to the population mean, simply because different samples contain different units. Sampling error is therefore an unavoidable feature of any sampling procedure. The good news is that, for a properly drawn random sample, the size of the sampling error can be calculated mathematically — this is the role of the standard error, which we shall meet in Section 2.13.

Two important properties of sampling error are worth committing to memory. First, sampling error decreases as the sample size n increases: a larger sample gives a more precise estimate, other things being equal. Second, sampling error decreases more slowly than the sample size increases: doubling the sample size does not halve the sampling error, it only reduces the sampling error by a factor of $\sqrt{2} \approx 1.41$. This square-root relationship has important practical consequences for survey design, as we shall see in Section 2.18.

2.11.2 Non-Sampling Errors

Non-sampling errors are all the other errors that can occur in a survey, quite apart from the fact that only a sample has been taken. They include errors of measurement (the respondent gives a wrong answer, perhaps because the question was poorly worded), errors of recording (the interviewer writes down the wrong number), errors of data entry (the data entry clerk presses the wrong key on the keyboard), errors of

classification (a respondent is wrongly assigned to one stratum instead of another), errors arising from non-response (some of the chosen units cannot be contacted, or refuse to answer), and errors of processing and computation (mistakes in the arithmetic or in the computer program). Unlike sampling errors, non-sampling errors do not necessarily diminish as the sample size grows; in fact, they often grow worse with larger samples, because larger samples involve larger field teams and more opportunities for human error.

Reducing non-sampling errors requires careful attention to all stages of the survey: well-designed questionnaires that have been pre-tested on a small group, well-trained interviewers, careful supervision of the field work, double-checked data entry, and skilled data analysts. In a well-conducted survey of moderate size, non-sampling errors are usually much larger than the sampling error itself. This is one of the most important practical reasons why a small, carefully conducted sample survey is often more accurate than a large, poorly conducted census.

2.12 THE SAMPLING DISTRIBUTION: A KEY CONCEPT

We now arrive at the single most important idea in this unit, and indeed in all of inferential statistics: the sampling distribution. The idea is initially a little subtle, so the student is encouraged to read this section slowly, perhaps more than once.

Suppose we draw a random sample of size n from some population, compute the sample mean $\bar{x}$, and write down the result. Now suppose we put the sample back into the population and draw a second random sample of the same size, compute its sample mean, and write that down too. And then a third sample, a fourth sample, a fifth sample, and so on. After we have done this many times — perhaps thousands, or even millions of times — we will have a long list of sample means, all from samples of the same size n drawn from the same population. The values in this long list will not all be the same, because different samples contain different units. They will, however, follow some pattern: most of the sample means will cluster around the true population mean μ, while only a few will be very far from μ. The pattern formed by all the possible sample means, together with their relative frequencies, is called the sampling distribution of the sample mean.

More generally, the sampling distribution of any sample statistic is the probability distribution of that statistic when computed for all possible samples of a given size from a given population. The sampling distribution is the bridge between the sample we actually have and the population we want to know about. It tells us how much sample-to-sample variability we should expect in our estimate, and it allows us to attach precise probabilities and confidence levels to our inferences. Without the concept of a sampling distribution, statistical inference would be impossible.

2.12.1 Two Important Things to Remember

First, the sampling distribution is a theoretical construct. We do not actually have to draw thousands of samples in order to use it; the mathematics of probability tells us in advance what the sampling distribution will look like, given the population and the sample size. In practice we draw exactly one sample, compute one statistic, and then use the theoretical sampling distribution to interpret what that one statistic tells us about the population.

Second, different statistics have different sampling distributions. The sampling distribution of the sample mean $\bar{x}$ is one thing; the sampling distribution of the sample proportion $\hat{p}$ is something else; and the sampling distribution of the sample variance s^2 is yet another thing. Each of these will be derived (or at least described) in the next several sections, with particular attention to the sample mean, which is by far the most important.

2.13 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

We now derive — or at least state — the sampling distribution of the sample mean $\bar{x}$. We begin with the simplest case, in which the population is small enough that we can actually enumerate all possible samples of size n. After we have grasped the basic idea in that simple case, we will state the general results that apply when the population is too large for direct enumeration.

2.13.1 A Small-Population Illustration

To make the idea concrete, let us start with a tiny population so that we can write down every single possible sample. Suppose the population consists of just N = 4 students, whose ages (in years) are 20, 22, 24 and 26. The population mean and the population variance can be computed directly:

$$\mu = \frac{20 + 22 + 24 + 26}{4} = 23\ years.$$

$$\sigma^2 = \frac{(20 - 23)^2 + (22 - 23)^2 + (24 - 23)^2 + (26 - 23)^2}{4} = \frac{9 + 1 + 1 + 9}{4} = 5,$$

$$\sigma = \sqrt{5} = 2.236\ years.$$

Now let us draw all possible samples of size n = 2 from this population, with replacement. There are N^n = 4^2 = 16 such samples, listed in the table below together with the sample mean of each:

Sample	Values	Sample Mean $\bar{x}$
1	(20, 20)	20
2	(20, 22)	21
3	(20, 24)	22
4	(20, 26)	23
5	(22, 20)	21
6	(22, 22)	22
7	(22, 24)	23
8	(22, 26)	24
9	(24, 20)	22
10	(24, 22)	23
11	(24, 24)	24
12	(24, 26)	25
13	(26, 20)	23
14	(26, 22)	24
15	(26, 24)	25
16	(26, 26)	26

The sample means take seven distinct values (20, 21, 22, 23, 24, 25, 26), with the following frequencies:

$\bar{x}$	20	21	22	23	24	25	26	Total
Frequency	1	2	3	4	3	2	1	16

$\bar{x}$	20	21	22	23	24	25	26	Total
Probability P($\bar{x}$)	1/16	2/16	3/16	4/16	3/16	2/16	1/16	1

This last table is the sampling distribution of the sample mean for our tiny example. Notice that it is symmetric and peaks at the centre, just at the value $\bar{x} = 23$ — which is exactly the population mean μ. We can compute the mean and variance of this sampling distribution directly:

$$\mu_{\bar{x}} = E(\bar{x}) = \frac{1(20) + 2(21) + 3(22) + 4(23) + 3(24) + 2(25) + 1(26)}{16} = \frac{368}{16} = 23.$$

And after some arithmetic (which the student should verify):

$$\sigma_{\bar{x}}^2 = \frac{40}{16} = 2.5, \quad \sigma_{\bar{x}} = \sqrt{2.5} = 1.581.$$

Two extremely important facts emerge from this small example. First, the mean of the sampling distribution of $\bar{x}$ is exactly equal to the population mean μ — that is, $\mu_{\bar{x}} = \mu$. Second, the variance of the sampling distribution of $\bar{x}$ is exactly equal to the population variance divided by the sample size — that is, $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$. In our example, $\sigma^2 = 5$ and $n = 2$, so $\sigma^2/n = 2.5$, exactly matching what we computed by enumeration. These two facts are not coincidences; they hold in general, as we now state formally.

2.13.2 General Results for the Sampling Distribution of $\bar{x}$

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population with mean μ and variance σ^2. Let $\bar{x}$ denote the sample mean. Then the sampling distribution of $\bar{x}$ has the following properties:

$$\mu_{\bar{x}} = E(\bar{x}) = \mu \qquad (2.4)$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \quad and \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (2.5)$$

The quantity $\sigma_{\bar{x}}$ in equation (2.5) is the standard deviation of the sampling distribution of $\bar{x}$ and is given a special name: it is called the standard error of the mean. The "standard error" is simply the standard deviation of a sampling distribution; the special name is used to distinguish it from the standard deviation of the original population (σ) and from the standard deviation computed from a single sample (s). The student should make absolutely sure that he or she understands the difference between these three closely related but distinct quantities, because confusing them is one of the commonest errors in introductory statistics.

The formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ holds exactly when sampling is from an infinite population, or equivalently when sampling is done with replacement from a finite population. When sampling is done without replacement from a finite population of size N, the formula must be multiplied by a finite population correction (FPC) factor:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad (2.6)$$

When the sample is small relative to the population — specifically, when n is less than about 5% of N — the FPC is very close to 1 and can usually be ignored. In nearly every social-science application this condition is satisfied, and we can use the simpler formula (2.5) without any correction.

Example 2.4 (Mean and Standard Error of $\bar{x}$)

A population has mean μ = 50 and standard deviation σ = 12. A simple random sample of size n = 36 is to be drawn from this population. Find the mean and the standard error of the sampling distribution of $\bar{x}$.

Solution. From equations (2.4) and (2.5):

$$\mu_{\bar{x}} = \mu = 50,$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2.$$

So the sampling distribution of the sample mean is centred at 50, with a standard error of 2. Notice how much smaller the standard error (2) is than the population standard deviation (12). This is the reduction in variability that we obtain by averaging — the central reason why the sample mean is a more reliable estimator of μ than any single observation could possibly be.

Example 2.4(b) (Standard Error from Sample Data)

A simple random sample of n = 64 final-year students at a public-sector university in Islamabad gave a mean monthly pocket money of $\bar{x}$ = 8,500 rupees with a sample standard deviation of s = 1,600 rupees. Estimate the standard error of the sample mean.

Solution. When the population standard deviation σ is unknown — as is almost always the case in real research — we estimate the standard error by replacing σ with the sample standard deviation s:

$$SE_{\bar{x}} \approx \frac{s}{\sqrt{n}} = \frac{1600}{\sqrt{64}} = \frac{1600}{8} = 200\ rupees.$$

So the standard error of the sample mean is about Rs. 200. This single number tells us how much the sample mean is likely to vary from one sample of 64 students to another, and it is the key ingredient in any confidence interval or hypothesis test we might wish to construct from these data.

Example 2.5 (Standard Error with the Finite Population Correction)

Suppose now that the population in Example 2.4 contains only N = 200 units, and the sample is drawn without replacement. Recompute the standard error of $\bar{x}$, applying the finite population correction.

Solution. Applying equation (2.6):

$$\sigma_{\bar{x}} = \frac{12}{\sqrt{36}} \cdot \sqrt{\frac{200 - 36}{200 - 1}} = 2 \cdot \sqrt{\frac{164}{199}} = 2 \cdot \sqrt{0.8241} = 2 \cdot 0.9078 = 1.816.$$

The corrected standard error is 1.816, slightly smaller than the uncorrected value of 2. The sample is 36/200 = 18% of the population, which is large enough that the correction is noticeable. If the same sample of 36 had been drawn from a much larger population — say N = 10,000 — the

correction factor would have been $\sqrt{(9964/9999)} \approx 0.9982$, which is so close to 1 that it could be ignored entirely.

2.14 THE CENTRAL LIMIT THEOREM

We have just learned that, for any population with mean μ and variance σ^2, the sampling distribution of the sample mean $\bar{x}$ has mean μ and variance σ^2/n. We have not yet said anything, however, about the shape of that sampling distribution. Is the sampling distribution normal? Skewed? Bell-shaped? Strangely irregular? It turns out that the answer depends on two things: the shape of the original population, and the size of the sample. The Central Limit Theorem (CLT) is the celebrated mathematical result that ties these two factors together. It is, without exaggeration, one of the most important theorems in all of statistics.

2.14.1 Statement of the Central Limit Theorem

The Central Limit Theorem can be stated in plain language as follows. If we draw a random sample of size n from a population with mean μ and finite standard deviation σ, then, as n becomes large, the sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean μ and standard error $\sigma/\sqrt{n}$, regardless of the shape of the original population. In symbols, for large n:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad (approximately, for\ large\ n) \qquad (2.7)$$

Equivalently, the standardized sample mean

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad (2.8)$$

approaches the standard normal distribution $N(0, 1)$ as n becomes large. This is the form in which the theorem is most often used in practice, because the Z statistic can be looked up directly in the standard normal table to compute probabilities and to construct confidence intervals.

2.14.2 What "Large n" Means in Practice

A natural question is: how large does n have to be for the Central Limit Theorem to give a good approximation? The honest answer is that it depends on the shape of the original population. If the population itself is already normal, the sampling distribution of $\bar{x}$ is exactly normal for every sample size, even for n as small as 2. If the population is approximately symmetric and bell-shaped, the sampling distribution of $\bar{x}$ is well approximated by a normal distribution for samples as small as $n \approx 10$ or 15. If the population is moderately skewed, a sample of about $n = 30$ is usually sufficient. If the population is extremely skewed or has very heavy tails, a much larger sample may be needed — perhaps $n = 100$ or more. The most commonly quoted rule of thumb in textbooks is that $n \geq 30$ is "usually large enough" for the Central Limit Theorem to apply, and we shall adopt this convention throughout the rest of this course.

2.14.3 Why the Central Limit Theorem Matters

The Central Limit Theorem is the single most important reason why the normal distribution occupies such a central place in statistics. It tells us that, no matter what shape the original population has, the sample mean of a sufficiently large sample is approximately normally distributed. This in turn means that the standard normal table — and the techniques built upon it, such as confidence intervals and tests of hypotheses — can be applied to a vast range of practical problems, even when we know nothing about the shape of the

population we are sampling from. Without the Central Limit Theorem, statistical inference about means would have to be done separately for every possible shape of population; with it, we can use a single, beautifully unified framework.

Example 2.6 (Using the Central Limit Theorem)

The monthly household expenditure in a large city in Pakistan has a mean of $\mu = 60$ thousand rupees and a standard deviation of $\sigma = 18$ thousand rupees. The shape of the population distribution is unknown but is believed to be moderately right-skewed. A simple random sample of $n = 100$ households is to be drawn. What is the (approximate) probability that the sample mean monthly expenditure exceeds 63 thousand rupees?

Solution. Although the shape of the population is unknown, the sample size $n = 100$ is well above 30, so the Central Limit Theorem applies and the sampling distribution of $\bar{x}$ is approximately normal with mean $\mu = 60$ and standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{100}} = \frac{18}{10} = 1.8 \; thousand \; rupees.$$

To find $P(\bar{x} > 63)$, we standardise:

$$Z = \frac{63 - 60}{1.8} = \frac{3}{1.8} = 1.667.$$

From the standard normal table, $P(Z > 1.67) \approx 0.0475$. So there is approximately a 4.75% chance that a random sample of 100 households will yield a sample mean exceeding 63 thousand rupees, even though the true population mean is only 60. This small probability is the kind of statement that the Central Limit Theorem allows us to make without knowing anything else about the shape of the population.

Example 2.6(b) (CLT — Probability Inside an Interval)

The mean weight of a particular variety of mango grown in Multan is $\mu = 250$ grams with a population standard deviation of $\sigma = 40$ grams. A wholesale buyer randomly selects a sample of $n = 64$ mangoes. Find the probability that the sample mean weight lies between 245 and 258 grams.

Solution. The sample size $n = 64$ is comfortably above 30, so by the Central Limit Theorem the sampling distribution of $\bar{x}$ is approximately normal with mean $\mu = 250$ grams and standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = \frac{40}{8} = 5 \; grams.$$

We standardise the lower and upper limits of the interval:

$$Z_{lower} = \frac{245 - 250}{5} = -1.00, \quad Z_{upper} = \frac{258 - 250}{5} = 1.60.$$

From the standard normal table:

$$P(245 \leq \bar{x} \leq 258) = P(-1.00 \leq Z \leq 1.60) = 0.9452 - 0.1587 = 0.7865.$$

There is therefore about a 78.65% chance that the sample mean weight of 64 randomly chosen mangoes will lie between 245 and 258 grams. The wholesale buyer can use this information for quality-control planning.

2.15 THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

Many of the most important questions in social-science research are not about means but about proportions. What proportion of voters intend to support a particular candidate? What proportion of households own a refrigerator? What proportion of students at a university are female? In each case the variable of interest is a yes/no attribute of each unit, and the parameter of interest is the proportion of units in the population that have that attribute. The same logic of sampling distributions that we developed for the sample mean can be applied to the sample proportion, with only minor modifications.

2.15.1 Setting Up the Problem

Let π denote the proportion of units in the population that possess some specified attribute (such as "owns a refrigerator", "intends to vote for candidate A", "passed the examination"). The complementary proportion $1 - \pi$ is the proportion of units in the population that do not possess the attribute. Now suppose we draw a random sample of size n from this population and count the number of units in the sample that possess the attribute. Call this count X. The sample proportion is then defined as:

$$\hat{p} = \frac{X}{n} \qquad (2.9)$$

$\hat{p}$ is the natural estimator of the population proportion π. If, for example, 240 out of 800 sampled voters say they intend to vote for candidate A, then $X = 240, n = 800$, and the sample proportion is $\hat{p} = 240/800 = 0.30$ — that is, 30% of the sampled voters intend to vote for the candidate.

2.15.2 Mean and Standard Error of the Sample Proportion

The sampling distribution of $\hat{p}$ has the following two basic properties. First, the mean of the sampling distribution of $\hat{p}$ is exactly equal to the population proportion π:

$$\mu_{\hat{p}} = E(\hat{p}) = \pi \qquad (2.10)$$

This is a very pleasant result: it tells us that the sample proportion is, on average, equal to the population proportion. We say that $\hat{p}$ is an unbiased estimator of π. Second, the standard deviation of the sampling distribution of $\hat{p}$ — that is, the standard error of the sample proportion — is given by:

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad (2.11)$$

Notice that this formula contains the unknown population proportion π. In practice, when we compute the standard error from sample data, we replace π with the sample proportion $\hat{p}$ and use the slightly approximate formula:

$$SE_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (2.12)$$

When sampling is from a finite population without replacement, the same finite population correction factor $\sqrt{[(N - n)/(N - 1)]}$ is applied to the standard error in (2.11), exactly as for the sample mean.

2.15.3 Normal Approximation to the Distribution of $\hat{p}$

The sample count X is the sum of n independent Bernoulli (0–1) random variables, so it follows a binomial distribution with parameters n and π. By the Central Limit Theorem, this binomial distribution is well approximated by a normal distribution when n is large enough. The conventional rule of thumb is that the normal approximation to the binomial — and therefore to the sampling distribution of $\hat{p}$ — is acceptable when both $n\pi \geq 5$ and $n(1 - \pi) \geq 5$. When this rule is satisfied:

$$\hat{p} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \quad (approximately) \qquad (2.13)$$

and the standardized sample proportion

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \qquad (2.14)$$

follows the standard normal distribution N(0, 1) approximately. This is the formula we shall use in Units 3 and 4 to construct confidence intervals and to test hypotheses about a population proportion.

Example 2.7 (Sampling Distribution of a Sample Proportion)

Past records suggest that 40% of the customers of a large supermarket in Karachi pay by debit card. A random sample of n = 150 customers will be observed on a particular day. (a) Find the mean and the standard error of the sampling distribution of the sample proportion $\hat{p}$ of debit-card payers. (b) Compute the (approximate) probability that the sample proportion will be between 0.35 and 0.45.

Solution. Here $\pi = 0.40$ and $n = 150$. The check $n\pi = 60 \geq 5$ and $n(1 - \pi) = 90 \geq 5$ confirms that the normal approximation is valid.

(a)

$$\mu_{\hat{p}} = \pi = 0.40,$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.40)(0.60)}{150}} = \sqrt{0.0016} = 0.04.$$

(b) We standardise the lower and upper limits:

$$Z_{lower} = \frac{0.35 - 0.40}{0.04} = \frac{-0.05}{0.04} = -1.25,$$

$$Z_{upper} = \frac{0.45 - 0.40}{0.04} = \frac{0.05}{0.04} = 1.25.$$

From the standard normal table:

$$P(0.35 \leq \hat{} \leq 0.45) = P(-1.25 \leq Z \leq 1.25) = 0.8944 - 0.1056 = 0.7888.$$

So there is approximately a 78.88% chance that the sample proportion of debit-card payers will fall between 0.35 and 0.45, even though the true proportion in the population of all customers is exactly 0.40.

2.16 THE RELATIONSHIP BETWEEN SAMPLE SIZE AND STANDARD ERROR

We close the unit with a closer look at the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ for the standard error of the sample mean. This formula has very important practical consequences for the design of any sample survey, and it deserves to be understood thoroughly. Three observations are particularly important.

2.16.1 The Standard Error Decreases as the Sample Size Increases

The most obvious thing about the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is that the standard error gets smaller as n gets bigger. This formalises the intuitive idea that a larger sample gives a more precise estimate. If we double the sample size, the precision improves; if we multiply the sample size by ten, the precision improves much more. Other things being equal, the only way to obtain a more precise estimate of μ is to use a larger sample.

2.16.2 But the Decrease is Slow — the Square Root Effect

The standard error contains the square root of n, not n itself. This means that doubling the sample size does not halve the standard error; it only divides it by $\sqrt{2} \approx 1.414$. To halve the standard error, we must multiply the sample size by 4. To divide the standard error by 10, we must multiply the sample size by 100. The following table makes this point vividly:

Sample size n	Standard error σ/√n	Reduction factor
25	σ/5 = 0.200 σ	1.00 (baseline)
100	σ/10 = 0.100 σ	0.50
400	σ/20 = 0.050 σ	0.25
2,500	σ/50 = 0.020 σ	0.10
10,000	σ/100 = 0.010 σ	0.05

Notice the dramatic implication: to reduce the standard error to one-tenth of its original value, we have to multiply the sample size by one hundred. This is one of the most important practical lessons of inferential statistics. A sample of 100 is dramatically more precise than a sample of 25, but a sample of 10,000 is only twice as precise as a sample of 2,500. There is, in other words, a law of diminishing returns: each additional unit of sample size buys less precision than the unit before it. For this reason, after a certain point it becomes much more efficient to invest the survey budget in reducing non-sampling errors — better questionnaires, better interviewers, better data processing — than in increasing the sample size still further.

2.16.3 The Standard Error Does Not Depend on the Population Size

A surprising and very important observation is that the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ does not contain N, the size of the population. In other words, what determines the precision of the sample mean is the size of the sample, not the size of the population. A sample of 1,000 households gives essentially the same precision whether the population has 50,000 households or 50 million households (provided in both cases the sample is small relative to the population, so that the finite-population correction can be ignored). This counter-intuitive fact is sometimes called the "soup-tasting principle": to taste whether a pot of soup is properly seasoned, you take a single small spoonful — and it does not matter at all whether the pot contains one litre or one hundred litres, as long as the soup has been well stirred. What matters is the size of the spoonful (the sample), not the size of the pot (the population).

Example 2.8 (a) (Sample Size and Precision)

A population has standard deviation σ = 50. (a) What sample size is needed to make the standard error of the sample mean equal to 5? (b) To make it equal to 2? (c) To make it equal to 1?

Solution. We use the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and solve for n in each case.

(a) We require σ/√n = 5, so √n = 50/5 = 10, and n = 100.

(b) We require σ/√n = 2, so √n = 50/2 = 25, and n = 625.

(c) We require σ/√n = 1, so √n = 50/1 = 50, and n = 2,500.

Reducing the standard error from 5 to 2 (a factor of 2.5) requires increasing the sample size from 100 to 625 (a factor of 6.25 = 2.5²). Reducing it from 5 to 1 (a factor of 5) requires increasing the sample size from 100 to 2,500 (a factor of 25 = 5²). The square-root relationship is unforgiving: high precision is expensive.

2.16.4 Practical Sample-Size Determination

The square-root relationship leads to a very practical formula for determining the sample size needed to achieve a desired precision in estimation. Suppose we want the margin of error of our 95% confidence interval for μ to be no larger than some specified value E. The half-width of the 95% confidence interval is $z_{0.025} \cdot \sigma/\sqrt{n} = 1.96 \cdot \sigma/\sqrt{n}$, so we require:

$$1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq E \quad \text{which gives} \quad n \geq \frac{(1.96\,\sigma)^2}{E^2} \qquad (2.16)$$

A similar formula holds for the sample proportion: to estimate π with a margin of error no larger than E at 95% confidence, the required sample size is $n \geq (1.96)^2\,\pi(1-\pi)\,/\,E^2$. If π is unknown in advance — as is often the case — the conservative choice π = 0.5 is used, because this maximizes the product π(1 − π) and so gives the largest possible (and therefore safest) sample size. We will return to sample-size determination in much more detail in Unit 3.

Example 2.8 (b) (Determining the Required Sample Size)

A researcher wishes to estimate the mean monthly expenditure on transportation by university students in Karachi to within Rs. 200 of the true value, with 95% confidence. From a small pilot study, the standard deviation of monthly transportation expenditure is estimated to be σ ≈ 1,400 rupees. What is the smallest sample size that will achieve this precision?

Solution. Applying equation (2.16) with σ = 1,400, E = 200 and $z_{0.025} = 1.96$:

$$n \geq \frac{(1.96 \times 1400)^2}{200^2} = \frac{(2744)^2}{40000} = \frac{7{,}529{,}536}{40{,}000} \approx 188.24.$$

Since n must be a whole number, we round up to the next integer: n = 189. The researcher should plan to sample at least 189 students. Notice the sensitivity of the answer to the desired precision: if the researcher had been willing to accept a margin of error of Rs. 400 instead of Rs. 200, the required

sample size would have been only $n \approx 47$ — about a quarter as many. Cutting the margin of error in half costs four times as many observations, exactly as the square-root relationship predicts.

2.17 SUMMARY

In this unit we have laid down the conceptual foundations for everything that will follow in the remaining units of the course. The main points may be summarized as follows.

- A population is the entire collection of units about which information is desired; a sample is a subset of the population that is actually examined. Quantities computed from the population are called parameters and are denoted by Greek letters (μ, σ^2, σ, π, ρ); quantities computed from a sample are called statistics and are denoted by Latin letters ($\bar{x}$, s^2, s, $\hat{p}$, r). Parameters are fixed but unknown; statistics are observable but vary from sample to sample.
- Sampling is preferred to complete enumeration in nearly every practical situation, because it is cheaper, faster, more accurate, more flexible and sometimes the only feasible method. The choice of sampling method depends on the nature of the population. Probability sampling methods (simple random, stratified, systematic, cluster, multi-stage) are the foundation of inferential statistics, while non-probability methods (convenience, quota, judgement, snowball) are used in qualitative and exploratory work.
- The total error in any sample survey consists of sampling error (which arises from observing only a sample) and non-sampling error (which arises from poor questionnaires, untrained interviewers, data-entry mistakes and so on). Sampling error can be calculated mathematically; non-sampling error must be controlled by good survey practice.
- The sampling distribution of a statistic is the probability distribution of that statistic over all possible samples of a given size from a given population. It is the bridge between the sample and the population, and it is the key concept on which all of inferential statistics rests.
- For the sample mean $\bar{x}$, the sampling distribution has mean μ and standard error $\sigma/\sqrt{n}$. The Central Limit Theorem tells us that, for sufficiently large n, this sampling distribution is approximately normal regardless of the shape of the population. The conventional rule of thumb is $n \geq 30$. For the sample proportion $\hat{p}$, the sampling distribution has mean π and standard error $\sqrt{[\pi(1-\pi)/n]}$, and the normal approximation is acceptable when $n\pi \geq 5$ and $n(1-\pi) \geq 5$.
- Statistical inference comes in two main forms: estimation (point and interval) and hypothesis testing. The general form of a confidence interval for a population mean, when σ is known, is $\bar{x} \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n}$. The width of the interval depends on the population variability, the sample size and the chosen confidence level.
- The standard error of the sample mean shrinks as $1/\sqrt{n}$. This square-root law has the practical consequence that high precision is expensive: to halve the standard error, the sample size must be multiplied by four. The size of the population, on the other hand, does not affect the standard error at all (provided the sample is small relative to the population).

In Unit 3 we shall develop estimation in much more detail, including the construction of confidence intervals for means, proportions, differences of means and variances, and the determination of the required sample size. In Units 4 and 5 we shall use exactly the same machinery — the sampling distribution and the Central Limit Theorem — to develop the theory of hypothesis testing, first for one sample and then for two samples.

The student who has mastered the material of the present unit will find these later units a natural extension of what he or she already knows; the student who has not will find them very difficult indeed.

2.18 SELF-ASSESSMENT EXERCISE

1. Distinguish clearly between a population and a sample, and between a parameter and a statistic. Explain why parameters are conventionally denoted by Greek letters and statistics by Latin letters.
2. List five practical advantages of sampling over a complete enumeration of the population. Give a real-world example for each advantage.
3. Distinguish between probability sampling and non-probability sampling. Name two methods of each kind and briefly describe one practical situation in which each would be the most appropriate.
4. Define the sampling distribution of a statistic. Why is it called the bridge between the sample and the population? What is the difference between the standard deviation and the standard error?
5. State the Central Limit Theorem in plain language. Explain why it occupies such a central place in inferential statistics, and discuss the conventional rule of thumb for the sample size required for the theorem to apply.
6. Discuss in detail the principal probability sampling methods — simple random, stratified, systematic, cluster and multi-stage sampling. For each method, explain the basic procedure, list the main advantages and disadvantages, and give a practical situation in which it would be the most appropriate choice. Compare and contrast stratified sampling and cluster sampling, paying particular attention to the difference in their objectives.
7. Develop the sampling distribution of the sample mean, beginning with the small-population case in which all possible samples can be enumerated, and moving to the general case described by the Central Limit Theorem. State the formulas for the mean and standard error of $\bar{x}$, including the finite population correction. Illustrate your discussion with at least one numerical example of your own choosing, and explain why the standard error decreases only as the square root of the sample size.
8. A population consists of N = 5 elements with values 4, 6, 8, 10 and 12. (a) Compute the population mean μ and population variance σ^2. (b) List all possible samples of size n = 2 drawn without replacement, and compute the sample mean for each. (c) Verify that the mean of the sampling distribution of $\bar{x}$ equals μ. [Answers: $\mu = 8$; $\sigma^2 = 8$; ten possible samples; mean of sampling distribution = 8]
9. A population has $\mu = 200$ and $\sigma = 30$. A simple random sample of n = 100 is drawn. (a) Find the mean and the standard error of the sampling distribution of $\bar{x}$. (b) Find the (approximate) probability that $\bar{x}$ exceeds 205. [Answers: $\mu_{\bar{x}} = 200$; $\sigma_{\bar{x}} = 3$; $P(\bar{x} > 205) \approx 0.0475$]
10. A sample of n = 49 households is drawn from a town in which the standard deviation of monthly electricity consumption is known to be $\sigma = 70$ units. The sample mean is $\bar{x} = 285$ units. Construct a 95% confidence interval for the mean monthly consumption μ in the town. [Answer: (265.4 ; 304.6) units]
11. In a random sample of 400 voters from a constituency in Karachi, 220 said they intended to vote for candidate A. (a) Compute the sample proportion $\hat{p}$ supporting candidate A. (b) Compute the standard error of $\hat{p}$. (c) Construct an approximate 95% confidence interval for the population proportion π. [Answers: $\hat{p} = 0.55$; SE ≈ 0.02488; CI $\approx (0.501 ; 0.599)$]

12. A college has N = 1,500 students enrolled in three departments: Department A (600 students), Department B (550 students) and Department C (350 students). A stratified sample of total size n = 150 is to be drawn using proportional allocation. How many students should be drawn from each department? [Answers: $n_A = 60$, $n_B = 55$, $n_C = 35$]

13. The lifetime of a certain brand of electric bulb has mean 1,200 hours and standard deviation 100 hours. A random sample of 64 bulbs is drawn. Find the (approximate) probability that the sample mean lifetime is between 1,180 and 1,220 hours. [Answer: ≈ 0.8904]

14. A population has $\sigma = 24$. What sample size is needed to make the standard error of the sample mean equal to (a) 4? (b) 2? (c) 1? [Answers: (a) n = 36; (b) n = 144; (c) n = 576]

15. A market researcher in Islamabad believes that 35% of households own a microwave oven. He plans to draw a random sample of n = 200 households. (a) Compute the standard error of the sample proportion $\hat{p}$ of households owning a microwave oven, assuming the true proportion is indeed 0.35. (b) Find the probability that the sample proportion will be less than 0.30. [Answers: (a) SE ≈ 0.0337; (b) P ≈ 0.0691]

16. A simple random sample of 36 employees at a multinational company in Lahore gave a mean monthly salary of $\bar{x} = 95{,}000$ rupees with a population standard deviation of $\sigma = 18{,}000$ rupees. Construct (a) a 90% and (b) a 99% confidence interval for the mean monthly salary μ of all employees of the company. [Answers: (a) (90,065 ; 99,935); (b) (87,260 ; 102,740)]

17. A sample of n = 80 students at AIOU is drawn to estimate the proportion who own a smartphone. Suppose 68 of them say they own one. Compute the sample proportion and construct an approximate 95% confidence interval for the population proportion of AIOU students who own a smartphone. [Answers: $\hat{p} = 0.85$; CI ≈ (0.772 ; 0.928)]

UNIT 3: ESTIMATION

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Distinguish clearly between an estimator and an estimate, and between point and interval estimation, and explain the role each plays in inferential statistics.
2. State and explain the four desirable properties of a good point estimator — unbiasedness, consistency, efficiency and sufficiency — and verify these properties for the standard estimators of the mean, the variance and the proportion.
3. Construct confidence intervals for a single population mean (with σ known and with σ unknown), for the difference between two means, for a single proportion, for the difference between two proportions, and for a population variance.
4. Use the standard normal table, the t-table and the chi-square table correctly, and explain when each is appropriate.
5. Determine the sample size required to estimate a population mean or proportion with a specified margin of error and a specified confidence level, and apply this skill to the design of practical research studies in the social sciences.

3.1 INTRODUCTION

In Unit 2 we developed the central concept of inferential statistics: the sampling distribution. We saw that, when a random sample is drawn from a population, the sample mean and the sample proportion are random variables with their own probability distributions, and that for sufficiently large samples those distributions are approximately normal by the Central Limit Theorem. The present unit takes the next logical step. Now that we have a sample in front of us, how do we use it to make a sensible guess about the value of the unknown population parameter? This is the question that the theory of estimation is designed to answer.

Estimation is one of the two main branches of statistical inference, the other being hypothesis testing (which we shall study in Units 4 and 5). The two branches are closely related and rest on the same theoretical foundation, but they answer different practical questions. Estimation asks: "what is the value of the unknown parameter?" Hypothesis testing asks: "is the value of the unknown parameter equal to some specified number?" In this unit we are concerned only with estimation; the closely related techniques of hypothesis testing will follow in the next two units.

Estimation itself comes in two forms, which we briefly met in Unit 2. In point estimation, the researcher gives a single numerical value as the best guess for the parameter — for example, "the mean monthly household income is estimated to be Rs. 64,250". In interval estimation, the researcher gives an entire range of plausible values together with a statement of how confident he or she is that the true parameter lies within that range — for example, "we are 95% confident that the mean monthly household income lies between Rs. 62,780 and Rs. 65,720". The interval estimate is much more informative than the point estimate alone, because it conveys not only the best guess but also the precision of that guess. For this reason interval estimation is more important in practice and will receive the bulk of our attention in this unit.

The unit is organized as follows. We begin in Sections 3.2 and 3.3 with the definition of a point estimator and its desirable properties. Section 3.4 outlines the standard methods by which good point estimators can be derived. Section 3.5 introduces the idea of an interval estimate and the closely related notion of a confidence interval. Sections 3.6 to 3.11 then take up the construction of confidence intervals for the most common parameters in social-science research: the population mean, the difference between two means, the population proportion, the difference between two proportions, and the population variance. Section 3.12 deals with the practical question of how large a sample is needed to achieve a specified precision. The unit closes with a summary and a self-assessment exercise.

The reader is strongly advised to keep Unit 2 close at hand while studying this unit, because almost every formula in what follows depends on the sampling distribution of some statistic — and the sampling distributions in question were derived in Unit 2. A student who has not fully digested the Central Limit Theorem will find the present unit quite difficult; one who has will find it a natural extension.

3.2 POINT ESTIMATION

We begin with the simpler of the two kinds of estimation. A point estimator of a population parameter θ is a sample statistic that is used as a single-number guess for the value of θ. The particular numerical value the statistic takes for a given sample is called the point estimate. By long-standing convention, the parameter being estimated is denoted by a lower-case Greek letter (most commonly θ), and the corresponding estimator is denoted by adding a "hat" symbol on top of the same letter, so that $\hat{\theta}$ denotes "the estimator of θ". When the formula is applied to actual sample data, the result is the point estimate.

A few examples will make the idea concrete. The sample mean $\bar{x}$ is the standard point estimator of the population mean μ; if a particular sample of 50 households gives $\bar{x}$ = 64,250 rupees, then 64,250 is the point

estimate of μ for that particular sample. The sample proportion $\hat{p}$ is the standard point estimator of the population proportion π; if 220 out of 400 sampled voters say they will vote for candidate A, then $\hat{p} = 0.55$ is the point estimate of π for that particular sample. The sample variance s^2 is the standard point estimator of the population variance σ^2. The sample correlation coefficient r is the standard point estimator of the population correlation ρ. Each of these estimators takes a different value for every sample drawn from the population — that is, each is itself a random variable, with its own sampling distribution, just as we discussed in Unit 2.

It is important to keep three terms straight, because students often confuse them.

1. Parameter (θ): a fixed but unknown numerical characteristic of the population (for example, μ).
2. Estimator ($\hat{\theta}$): a rule or formula, defined in terms of the sample observations, that produces a guess for the parameter (for example, $\bar{x} = (1/n)\,\Sigma x_i$).
3. Estimate: the particular numerical value that the estimator produces for a given sample (for example, $\bar{x} = 64{,}250$).

The estimator is the formula; the estimate is the number that comes out when the formula is applied to data. Different samples produce different estimates from the same estimator. The reliability of an estimator depends not on any single estimate it happens to produce, but on the average behavior of the estimator across all possible samples — that is, on its sampling distribution.

3.3 PROPERTIES OF A GOOD POINT ESTIMATOR

Many different formulas can be proposed as estimators of the same parameter. For example, to estimate the central value of a population we could use the sample mean, the sample median, the sample mid-range (the average of the largest and smallest observations), or any number of other quantities. How do we choose between them? The theory of estimation answers this question by laying down certain desirable properties that a good estimator should possess. Four such properties are particularly important: unbiasedness, consistency, efficiency and sufficiency. We discuss each in turn.

3.3.1 Unbiasedness

An estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if its expected value (the long-run average over all possible samples) is exactly equal to θ:

$$E(\hat{\theta}) = \theta \qquad (3.1)$$

In words, an unbiased estimator is one that does not systematically over-estimate or under-estimate the true parameter. If we were to draw all possible samples of size n from the population, compute the estimate for each, and then average those estimates, the average would be exactly θ. Unbiasedness is therefore a guarantee that the estimator is "centred at the right place" — even though any individual estimate may differ from θ because of sampling error, the estimator as a whole has no systematic tendency to err in either direction.

We have already seen, in Unit 2, that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ, because we proved that $E(\bar{x}) = \mu$. Similarly, the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion π, because $E(\hat{p}) = \pi$. The sample variance s^2 (computed with the divisor $n - 1$) is an unbiased estimator of the population variance σ^2. It is precisely to obtain this unbiasedness property that we use the divisor $(n - 1)$ rather than n in the sample variance formula; if we used n instead, the resulting "biased sample variance" would systematically under-estimate σ^2. This is the reason behind the $(n - 1)$ divisor that we adopted in Unit 1.

3.3.2 Consistency

An estimator $\hat{\theta}$ is said to be a consistent estimator of θ if it converges (in probability) to θ as the sample size n becomes large. Loosely speaking, this means that, by taking a large enough sample, we can make the estimate as close to the true value as we like. More formally, for any small positive number ε, the probability that $|\hat{\theta} - \theta|$ is less than ε approaches 1 as n approaches infinity:

$$P(|\hat{\theta} - \theta| < \varepsilon) \to 1 \quad as \quad n \to \infty \qquad (3.2)$$

Consistency is a less stringent requirement than unbiasedness, because a consistent estimator is allowed to be biased for finite n, provided the bias disappears as n grows. Most of the standard estimators we work with, however, are both unbiased and consistent. The sample mean $\bar{x}$, the sample proportion $\hat{p}$, and the sample variance s^2 are all consistent estimators of their respective population parameters. Consistency is what justifies our intuitive belief that "more data is always better": with a consistent estimator, a larger sample size gives a more accurate estimate, in a sense that can be made mathematically precise.

3.3.3 Efficiency

Among the unbiased estimators of a given parameter, the one with the smallest variance (the smallest standard error) is called the most efficient estimator, or the minimum-variance unbiased estimator. The reason is simple: a smaller variance means that the estimator is more tightly clustered around the true parameter value across different samples, so the estimate produced by any single sample is more likely to be close to the truth.

A useful illustration is the comparison between the sample mean and the sample median as estimators of the centre of a normally distributed population. Both are unbiased, but it can be shown mathematically that the variance of the sample mean is smaller than the variance of the sample median by a factor of about 0.637. In other words, when sampling from a normal population, the sample mean is about 1.57 times more efficient than the sample median: a sample mean computed from 100 observations carries roughly the same information as a sample median computed from 157 observations. This is one of the reasons why the mean is so much more widely used than the median in inferential statistics, in spite of the median's well-known robustness against outliers.

3.3.4 Sufficiency

A statistic θ is said to be sufficient for the parameter θ if it captures all the information about θ that the sample contains. More precisely, a sufficient statistic is one such that the conditional distribution of the sample, given the value of the statistic, does not depend on θ. The intuitive idea is that a sufficient statistic "uses every drop" of the information in the sample about the parameter — knowing the sample in addition to the statistic adds nothing further about θ. The sample mean $\bar{x}$ is a sufficient statistic for the mean μ of a normal population with known variance, and the sample sum is sufficient for the rate parameter of a Poisson population. Sufficiency is a rather technical concept, and the student need not understand its formal proof at this stage; it is enough to remember that the standard estimators ($\bar{x}$, $\hat{p}$, s^2) are all sufficient for the parameters they estimate, in the populations most commonly encountered in social-science work.

Example 3.1 (Verifying Unbiasedness of the Sample Mean)

A small population consists of N = 4 observations: 2, 4, 6, 8. List all possible samples of size n = 2 (with replacement), compute the sample mean for each, and verify that $E(\bar{x}) = \mu$.

Solution. The population mean is $\mu = (2 + 4 + 6 + 8)/4 = 5$. The $4^2 = 16$ possible samples and their means are:

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
(2,2)	2	(4,2)	3	(6,2)	4	(8,2)	5
(2,4)	3	(4,4)	4	(6,4)	5	(8,4)	6
(2,6)	4	(4,6)	5	(6,6)	6	(8,6)	7
(2,8)	5	(4,8)	6	(6,8)	7	(8,8)	8

Summing all 16 sample means: $2 + 3 + 4 + 5 + 3 + 4 + 5 + 6 + 4 + 5 + 6 + 7 + 5 + 6 + 7 + 8 = 80$. The expected value of $\bar{x}$ is therefore:

$$E(\bar{x}) = \frac{80}{16} = 5 = \mu.$$

The mean of the 16 possible sample means is exactly equal to the population mean. This numerically verifies that $\bar{x}$ is an unbiased estimator of μ — a result we proved in general in Unit 2 and which holds for any population, no matter how its values are distributed.

3.4 METHODS OF OBTAINING POINT ESTIMATORS

Where do estimators actually come from? A natural question, since we have so far been treating $\bar{x}$, $\hat{p}$ and s^2 as if they had simply fallen from the sky. In fact, there are several systematic methods by which good estimators can be derived from first principles. Three of the most important methods are the method of moments, the method of maximum likelihood, and the method of least squares. We shall describe each of them only briefly here, leaving the more technical mathematical details to specialized texts.

3.4.1 The Method of Moments

The method of moments, due to the English statistician Karl Pearson, is the oldest and conceptually simplest of the three methods. The idea is to set the sample moments equal to the corresponding population moments and solve for the unknown parameters. For example, if a population has unknown mean μ, the population first moment is μ itself and the sample first moment is $\bar{x}$. Setting these equal gives $\hat{\mu} = \bar{x}$ — that is, the method-of-moments estimator of μ is simply the sample mean. If the population has both an unknown mean μ and an unknown variance σ^2, we use the first two sample moments: setting the first sample moment equal to μ and the second sample central moment equal to σ^2 gives $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = (1/n)\,\Sigma(x_i - \bar{x})^2$. The method of moments is intuitive and easy to apply, but the resulting estimators are not always the most efficient.

3.4.2 The Method of Maximum Likelihood

The method of maximum likelihood, developed by Sir Ronald A. Fisher in the 1920s, is the most widely used method of estimation in modern statistics. The idea is to find the value of the parameter that makes the observed sample most likely to have occurred. More precisely, we write down the joint probability of the sample as a function of the unknown parameter θ — this function, viewed as a function of θ rather than of

the data, is called the likelihood function L(θ) — and we then choose as our estimate the value of θ that maximizes L(θ). The resulting estimate is called the maximum likelihood estimate (MLE) of θ.

To take a simple example, suppose we draw a random sample of size n from a normal population with unknown mean μ and known variance σ^2. The likelihood function (after some simplification) is maximized when $\hat{\mu} = \bar{x}$, so the MLE of μ is the sample mean. Similarly, the MLE of the proportion π in a Bernoulli population is the sample proportion $\hat{p}$. Maximum likelihood estimators have several desirable properties: they are consistent, they are asymptotically unbiased, they are asymptotically normal, and they are asymptotically the most efficient possible. For these reasons MLE is the dominant method in modern inferential statistics, although the formal derivation of MLEs sometimes requires considerable mathematical work.

3.4.3 The Method of Least Squares

The method of least squares, developed by Carl Friedrich Gauss and Adrien-Marie Legendre in the early nineteenth century, chooses as its estimate the value that minimizes the sum of squared deviations between the observed data and what the model would predict. For the simple problem of estimating a population mean from a sample, least squares gives $\hat{\mu} = \bar{x}$, exactly as the other two methods do. The method of least squares becomes much more important — and much more interesting — in the context of regression analysis, where it is used to fit a straight line (or a more complicated curve) to bivariate or multivariate data. We shall develop the method of least squares in detail in Unit 7 on Regression and Correlation Analysis.

In the present unit we shall not derive estimators from first principles. Instead, we shall simply use the standard estimators that all three of the above methods produce — the sample mean, the sample proportion and the sample variance — and concentrate on how to use them to construct confidence intervals.

3.5 INTERVAL ESTIMATION AND CONFIDENCE INTERVALS

A point estimate, however carefully constructed, has one important weakness: it gives no indication of how reliable it is. If a researcher reports that the mean monthly household income in District Rawalpindi is estimated to be Rs. 64,250 on the basis of a sample of 600 households, the listener has no way of knowing whether the true mean is likely to be very close to that figure (say, between Rs. 63,500 and Rs. 65,000) or considerably further away (say, anywhere between Rs. 55,000 and Rs. 73,000). To convey not only the best guess but also its precision, we need an interval estimate, also called a confidence interval.

A confidence interval is a range of values, computed from sample data, that is likely to contain the true value of the unknown population parameter. Together with the interval, the researcher reports a confidence level — a number such as 90%, 95% or 99% — which describes how confident he or she is that the interval contains the true parameter. The most common confidence level by far is 95%, which has become almost a default in social-science research. The choice of confidence level depends on how serious the consequences of a wrong inference would be: in routine work 95% is usually sufficient, but in fields where the cost of an error is very high (such as medical research or quality control of pharmaceuticals) confidence levels of 99% or even 99.9% may be used.

3.5.1 The Anatomy of a Confidence Interval

Almost every confidence interval that we shall meet in this unit takes the same general form:

$$Estimate \pm (Critical\ Value) \times (Standard\ Error) \qquad (3.3)$$

The three ingredients of this formula are as follows. The estimate is the point estimate of the parameter — that is, the value computed from the sample by some standard estimator ($\bar{x}$, $\hat{p}$, s^2, and so on). The critical value is a number drawn from a probability table (the standard normal table, the t-table or the chi-square table, depending on the situation) that depends on the chosen confidence level. The standard error is the standard deviation of the sampling distribution of the estimator, as we developed it in Unit 2 ($\sigma/\sqrt{n}$ for the sample mean, $\sqrt{\pi(1-\pi)/n}$ for the sample proportion, and so on). The product of the critical value and the standard error is called the margin of error or the maximum likely error of the estimate; it is the half-width of the confidence interval.

3.5.2 Interpreting a Confidence Interval

The interpretation of a confidence interval is one of the trickiest issues in introductory statistics, and many students get it wrong on their first attempt. The correct interpretation is the following. If we were to draw many independent samples of the same size from the same population, and construct a 95% confidence interval from each sample using the same procedure, then in the long run about 95% of those intervals would contain the true value of the parameter, and about 5% would not. The "95% confidence" refers to the long-run reliability of the procedure, not to the probability that the true parameter lies in any one particular interval.

It is incorrect, strictly speaking, to say "there is a 95% probability that the true mean lies in the interval (305.3 ; 334.7)". The true mean is a fixed (though unknown) number; either it lies in this particular interval or it does not, and there is no probability about it. What we can say is that the procedure used to construct the interval has a 95% long-run success rate. In practice, however, the looser phrasing "we are 95% confident that the true mean lies between 305.3 and 334.7" is so widely used and so well understood that no harm comes from using it, provided we keep in mind what is really being said.

3.6 CONFIDENCE INTERVAL FOR THE POPULATION MEAN (σ KNOWN)

We begin with the simplest case: estimating the population mean μ when the population standard deviation σ is known. This case is somewhat artificial in practice — if we know σ, we very often know μ as well — but it is the natural starting point because the formulas are clean and the underlying logic is easiest to see. We shall move on in the next section to the more realistic case in which σ is unknown and must be estimated from the sample.

3.6.1 Derivation of the Formula

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population with unknown mean μ and known standard deviation σ. From Unit 2 we know that, for sufficiently large n (or for any n if the population itself is normal), the sampling distribution of the sample mean $\bar{x}$ is approximately:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardising $\bar{x}$, the quantity $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ follows the standard normal distribution N(0, 1). For a 95% confidence interval, we want to find the value of Z that cuts off 2.5% in each tail of the standard normal distribution. From the standard normal table, this critical value is $Z_{0.025} = 1.96$. We therefore have:

$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

Rearranging the inequality inside the brackets to isolate μ:

$$P\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

This is the 95% confidence interval for μ when σ is known. The general form, for any confidence level $(1 - \alpha) \times 100\%$, is:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \qquad (3.4)$$

where $Z_{\alpha/2}$ is the value from the standard normal table corresponding to the chosen confidence level. The most commonly used critical values are summarized in the following table:

Confidence Level	α	α/2	Critical Value $Z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
98%	0.02	0.01	2.33
99%	0.01	0.005	2.58
99.9%	0.001	0.0005	3.29

The student should commit at least the first three rows of this table to memory; they will be needed again and again in the rest of the course. The values 1.645 (90%), 1.96 (95%) and 2.58 (99%) are among the most frequently used numbers in all of applied statistics.

Example 3.2 (95% Confidence Interval for μ, σ Known)

A random sample of n = 64 students at a university in Lahore gave a mean monthly pocket money of $\bar{x}$ = 9,200 rupees. From past studies, the population standard deviation of monthly pocket money among university students in Lahore is known to be σ = 1,600 rupees. Construct a 95% confidence interval for the mean monthly pocket money μ of all students at this university.

Solution. We are given $\bar{x}$ = 9,200, σ = 1,600 and n = 64. For a 95% confidence interval, the critical value is $z_{0.025}$ = 1.96. Applying equation (3.4):

$$CI = 9200 \pm 1.96 \cdot \frac{1600}{\sqrt{64}} = 9200 \pm 1.96 \cdot \frac{1600}{8} = 9200 \pm 1.96 \cdot 200$$
$$= 9200 \pm 392.$$

So the 95% confidence interval for μ is (8,808 ; 9,592) rupees. We can therefore state with 95% confidence that the true mean monthly pocket money of all students at this university lies between approximately Rs. 8,800 and Rs. 9,600. The margin of error is Rs. 392, which represents about 4.3% of the point estimate — a reasonably precise interval for a sample of 64.

Example 3.3 (90%, 95% and 99% Intervals — Comparison)

Using the same sample as in Example 3.2 ($\bar{x}$ = 9,200, σ = 1,600, n = 64), construct 90%, 95% and 99% confidence intervals for μ. Compare their widths and comment on the trade-off between confidence and precision.

Solution. The standard error is σ/√n = 1,600/8 = 200, the same in all three cases. Only the critical value changes:

36. 90% CI: 9,200 ± 1.645 × 200 = 9,200 ± 329 = (8,871 ; 9,529). Width = 658.

37. 95% CI: 9,200 ± 1.96 × 200 = 9,200 ± 392 = (8,808 ; 9,592). Width = 784.

38. 99% CI: 9,200 ± 2.58 × 200 = 9,200 ± 516 = (8,684 ; 9,716). Width = 1,032.

As the confidence level increases from 90% to 99%, the width of the interval increases from 658 to 1,032 — that is, by about 57%. This illustrates the fundamental trade-off in interval estimation: greater confidence comes at the cost of a wider, less precise interval. The only way to obtain both higher confidence and a narrower interval simultaneously is to increase the sample size.

3.7 CONFIDENCE INTERVAL FOR THE POPULATION MEAN (σ UNKNOWN): THE t-DISTRIBUTION

In nearly every real-world investigation, the population standard deviation σ is just as unknown as the population mean μ. The natural way of getting around this difficulty is to estimate σ by the sample standard deviation s, computed from the same sample as $\bar{x}$, and to plug s into the formula in place of σ. When we do this, however, the resulting standardized quantity

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

no longer follows the standard normal distribution exactly, because s itself is a random variable that varies from sample to sample. Instead it follows a closely related but slightly different distribution called the Student's t-distribution, named after the English statistician William Sealy Gosset who published under the pen-name "Student" in 1908.

3.7.1 Properties of the t-Distribution

The t-distribution looks very similar to the standard normal distribution: it is symmetric about zero, bell-shaped, and centred at the same place. It differs from the standard normal in that its tails are slightly heavier — that is, more probability is concentrated in the extremes — to allow for the additional uncertainty introduced by estimating σ. The exact shape of the t-distribution depends on a parameter called the degrees of freedom, denoted by ν (the Greek letter nu) or df. For the case of estimating a single population mean from a sample of size n, the degrees of freedom are df = n − 1, the same divisor we used in the sample variance formula. As n grows large, the t-distribution approaches the standard normal more and more closely; for n greater than about 30 the two are practically indistinguishable.

3.7.2 Confidence Interval Formula

The (1 − α) × 100% confidence interval for μ when σ is unknown takes the same general form as before, but with $Z_{\alpha/2}$ replaced by the corresponding t-value and σ replaced by s:

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \qquad (3.5)$$

where $t_{\alpha/2,n-1}$ is the value from the t-table with (n − 1) degrees of freedom that cuts off α/2 in each tail. For example, for a 95% confidence interval based on a sample of size n = 20, the degrees of freedom are 19, and from the t-table the critical value is $t_{0.025,19} = 2.093$. For larger samples (say n ≥ 30), the t-value is so close

to the corresponding z-vlaue that it makes very little practical difference which one is used; some textbooks therefore recommend using the standard normal critical value as an approximation when n is large, and using the t-value strictly only for smaller samples. In this course we shall always use the t-value when σ is unknown, regardless of sample size, because that is the technically correct procedure.

Example 3.4 (Confidence Interval for μ, σ Unknown, Small Sample)

A random sample of n = 16 newly graduated MBA students from a business school in Karachi gave a mean starting monthly salary of $\bar{x}$ = 85,000 rupees with a sample standard deviation of s = 12,000 rupees. Assuming starting salaries are approximately normally distributed, construct a 95% confidence interval for the mean starting monthly salary μ of all newly graduated MBA students from this school.

Solution. Here $\bar{x}$ = 85,000, s = 12,000, and n = 16. The degrees of freedom are df = n − 1 = 15. For a 95% confidence interval, we need $t_{0.025,15}$. From the t-table, this value is 2.131. Applying equation (3.5):

$$CI = 85000 \pm 2.131 \cdot \frac{12000}{\sqrt{16}} = 85000 \pm 2.131 \cdot \frac{12000}{4} = 85000 \pm 2.131 \cdot 3000$$
$$= 85000 \pm 6393.$$

So the 95% confidence interval for μ is approximately (78,607 ; 91,393) rupees. We can state with 95% confidence that the mean starting monthly salary of all newly graduated MBA students from this school lies between approximately Rs. 78,600 and Rs. 91,400. Notice that, because the sample is small (n = 16), the t-critical value 2.131 is noticeably larger than the corresponding z-critical value 1.96; the resulting confidence interval is therefore slightly wider than it would have been if σ had been known. This widening reflects the additional uncertainty caused by estimating σ from such a small sample.

Example 3.5 (Confidence Interval for μ, σ Unknown, Larger Sample)

A random sample of n = 50 households in a small town near Faisalabad gave a mean monthly expenditure on food of $\bar{x}$ = 18,500 rupees with a sample standard deviation of s = 4,200 rupees. Construct a 99% confidence interval for the mean monthly food expenditure μ of all households in the town.

Solution. Here $\bar{x}$ = 18,500, s = 4,200, and n = 50, so the degrees of freedom are df = 49. For a 99% confidence interval, we need $t_{0.005,49}$. From the t-table, this value is approximately 2.680. Applying equation (3.5):

$$CI = 18500 \pm 2.680 \cdot \frac{4200}{\sqrt{50}} = 18500 \pm 2.680 \cdot \frac{4200}{7.071} = 18500 \pm 2.680 \cdot 593.97$$
$$= 18500 \pm 1591.84.$$

So the 99% confidence interval for μ is approximately (16,908 ; 20,092) rupees. We can state with 99% confidence that the true mean monthly food expenditure of all households in this town lies between approximately Rs. 16,900 and Rs. 20,100. Notice that with n = 50, the t-critical value 2.680 is only slightly larger than the corresponding z-critical value 2.576; the difference between using t

and z would have been only about 4% in this case, illustrating that for moderately large samples the two are almost interchangeable.

3.8 CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO MEANS

In many social-science studies the researcher is interested not in a single mean but in the comparison of two means. Is the average income in urban areas higher than in rural areas? Do male students score better than female students on a particular test? Does a new teaching method produce higher learning gains than the traditional method? In each of these cases the parameter of interest is the difference between two population means, $\mu_1 - \mu_2$, and we wish to construct a confidence interval for this difference.

The procedure parallels what we did for a single mean: we use the sample analogue $(\bar{x}_1 - \bar{x}_2)$ as a point estimate of $\mu_1 - \mu_2$, and we add and subtract the appropriate critical value times the standard error. The only complication is that the formula for the standard error depends on whether we know the population variances and on whether we are willing to assume that the two populations have the same variance.

3.8.1 Independent Samples, Population Variances Known

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of two independent random samples of sizes n_1 and n_2 drawn from two populations with known variances σ_1^2 and σ_2^2. It can be shown — using the same logic as in Section 2.13 — that the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ has mean $(\mu_1 - \mu_2)$ and standard error:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (3.6)$$

and is approximately normal for sufficiently large n_1 and n_2 (or exactly normal if both populations are themselves normal). The $(1 - \alpha) \times 100\%$ confidence interval for $\mu_1 - \mu_2$ is therefore:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (3.7)$$

3.8.2 Independent Samples, Population Variances Unknown But Assumed Equal

More commonly, the population variances are unknown and must be estimated from the samples. If we are willing to assume that the two populations have the same (unknown) variance, denoted by σ^2, we can estimate σ^2 by the pooled sample variance, which combines the information in both samples:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (3.8)$$

and the resulting confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (3.9)$$

with degrees of freedom $df = n_1 + n_2 - 2$. This is the so-called pooled-variance t-interval. It is the standard procedure when both samples come from approximately normal populations and the assumption of equal variances is reasonable.

3.8.3 Independent Samples, Population Variances Unknown and Unequal

When the assumption of equal variances is doubtful, an alternative procedure (known as Welch's approximation) is available. The formula for the confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (3.10)$$

with degrees of freedom computed by a slightly complicated formula due to Welch (which we shall not write out here). For the purposes of this introductory course, the student is encouraged to use the pooled-variance procedure (equation 3.9) whenever the two sample standard deviations are roughly comparable in size, and the unpooled procedure (equation 3.10) only when they differ markedly.

Example 3.6 (CI for the Difference Between Two Means, Pooled t)

A study was conducted to compare the mean monthly food expenditure of urban and rural households in Punjab. A random sample of $n_1 = 25$ urban households gave a mean of $\bar{x}_1 = 22{,}400$ rupees with a sample standard deviation of $s_1 = 5{,}200$ rupees, while an independent random sample of $n_2 = 30$ rural households gave a mean of $\bar{x}_2 = 16{,}800$ rupees with a sample standard deviation of $s_2 = 4{,}500$ rupees. Assuming the two populations have equal variances, construct a 95% confidence interval for the difference $(\mu_1 - \mu_2)$ in mean monthly food expenditure between urban and rural households.

Solution. We first compute the pooled sample variance, using equation (3.8):

$$s_p^2 = \frac{(25 - 1)(5200^2) + (30 - 1)(4500^2)}{25 + 30 - 2} = \frac{24(27{,}040{,}000) + 29(20{,}250{,}000)}{53}$$

$$= \frac{648{,}960{,}000 + 587{,}250{,}000}{53} = \frac{1{,}236{,}210{,}000}{53} = 23{,}324{,}716.98.$$

So $s_p = \sqrt{23{,}324{,}716.98} \approx 4{,}829.57$ rupees. The degrees of freedom are $df = n_1 + n_2 - 2 = 25 + 30 - 2 = 53$. For a 95% confidence interval, $t_{0.025,53} \approx 2.006$. Applying equation (3.9):

$$CI = (22400 - 16800) \pm 2.006 \cdot 4829.57 \cdot \sqrt{\frac{1}{25} + \frac{1}{30}}$$

$$= 5600 \pm 2.006 \cdot 4829.57 \cdot \sqrt{0.0733} = 5600 \pm 2.006 \cdot 4829.57 \cdot 0.2708$$
$$= 5600 \pm 2624.$$

The 95% confidence interval for $(\mu_1 - \mu_2)$ is therefore approximately (2,976 ; 8,224) rupees. We can state with 95% confidence that urban households in Punjab spend, on average, between Rs. 2,976 and Rs. 8,224 more per month on food than rural households. Notice that the entire interval lies above zero, which is strong evidence that urban food expenditure really is higher than rural food expenditure (an idea we shall develop more formally in Unit 5 on hypothesis testing).

3.9 CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Many of the most important parameters in social-science research are not means but proportions. What proportion of voters support a particular candidate? What proportion of households own a refrigerator? What proportion of patients are cured by a new treatment? In each of these cases the variable of interest is a yes/no attribute, and the parameter of interest is the population proportion π. We saw in Section 2.15 of Unit 2 that

the sampling distribution of the sample proportion $\hat{p}$ is approximately normal — by the Central Limit Theorem — when both $n\pi \geq 5$ and $n(1-\pi) \geq 5$, with mean π and standard error $\sqrt{[\pi(1-\pi)/n]}$.

The same logic that gave us the confidence interval for a mean gives us the confidence interval for a proportion. The $(1-\alpha) \times 100\%$ confidence interval for π takes the form "estimate ± critical value × standard error":

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (3.11)$$

Notice that, in computing the standard error, we have replaced the unknown population proportion π with the sample proportion $\hat{p}$. This substitution is necessary because we do not know π in advance, and it is acceptable provided n is large enough that $\hat{p}$ is a reasonably accurate estimate of π. The conventional rule of thumb is that the normal approximation is acceptable when both $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$; some textbooks use the slightly more conservative requirement $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$. Either rule is satisfied in nearly all practical applications, since most surveys have sample sizes of several hundred at least.

Example 3.7 (Confidence Interval for a Population Proportion)

In a pre-election survey conducted in a constituency in Karachi, a random sample of n = 400 voters was asked which candidate they intended to vote for. A total of 224 said they would vote for candidate A. (a) Compute the sample proportion supporting candidate A. (b) Construct a 95% confidence interval for the true population proportion π of all voters in the constituency who intend to support candidate A. (c) On the basis of this interval, can we conclude that candidate A has the support of more than half of the voters?

Solution.

(a) The sample proportion is $\hat{p} = 224/400 = 0.56$.

(b) Both $n\hat{p} = 224$ and $n(1-\hat{p}) = 176$ are well above 5, so the normal approximation is valid. The standard error is:

$$SE = \sqrt{\frac{(0.56)(0.44)}{400}} = \sqrt{0.000616} = 0.02482.$$

For a 95% confidence interval, $z_{0.025} = 1.96$, and the margin of error is $1.96 \times 0.02482 = 0.0487$. Applying equation (3.11):

$$CI = 0.56 \pm 0.0487 = (0.5113\,;\, 0.6087).$$

(c) The entire confidence interval lies above 0.50. We can therefore state with 95% confidence that more than half of the voters in this constituency support candidate A. This is exactly the kind of inference that pre-election polls are meant to support, and it is a more honest report than simply quoting the point estimate of 56% without any indication of its precision.

Example 3.8 (CI for a Proportion — Larger Sample)

A national consumer-protection agency wants to estimate the proportion of households in Pakistan that own a smartphone. A random sample of n = 1,500 households is drawn, and 1,170 are found

to own at least one smartphone. Construct a 99% confidence interval for the population proportion π.

Solution. The sample proportion is $\hat{p} = 1{,}170/1{,}500 = 0.78$. The standard error is:

$$SE = \sqrt{\frac{(0.78)(0.22)}{1500}} = \sqrt{0.0001144} = 0.01070.$$

For a 99% confidence interval, $z_{0.005} = 2.576$, and the margin of error is $2.576 \times 0.01070 = 0.02756$. The 99% confidence interval is:

$$CI = 0.78 \pm 0.02756 = (0.7524\,;\, 0.8076).$$

We can state with 99% confidence that the proportion of households in Pakistan owning a smartphone lies between approximately 75.2% and 80.8%. Notice that the margin of error here (about 2.8%) is much smaller than in Example 3.7 (about 4.9%), even though the confidence level here is higher. The reason is the much larger sample size (1,500 versus 400) — once again confirming the principle that precision is bought with sample size.

Example 3.8(b) (CI for a Proportion — Education Survey)

A study of educational attainment in a rural district of Sindh draws a random sample of n = 600 adults aged 25 years or older and finds that 162 of them have completed at least a secondary-school education. (a) Compute the sample proportion. (b) Construct a 95% confidence interval for the population proportion π of adults in this district who have completed at least a secondary-school education. (c) Comment on what the interval suggests about educational attainment in the district.

Solution.

(a) The sample proportion is $\hat{p} = 162/600 = 0.27$.

(b) Both $n\hat{p} = 162$ and $n(1 - \hat{p}) = 438$ are well above 5, so the normal approximation is valid. The standard error is:

$$SE = \sqrt{\frac{(0.27)(0.73)}{600}} = \sqrt{0.0003285} = 0.01813.$$

For a 95% confidence interval, $z_{0.025} = 1.96$, and the margin of error is $1.96 \times 0.01813 = 0.03554$. Applying equation (3.11):

$$CI = 0.27 \pm 0.03554 = (0.2345\,;\, 0.3055).$$

(c) We can state with 95% confidence that between approximately 23.5% and 30.5% of adults aged 25 or older in this district have completed at least secondary-school education. Notice that the entire interval lies well below 0.50, which suggests that fewer than half of all adults in the district have reached this educational level — a finding of clear policy relevance for the provincial government.

3.10 CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

Just as we may wish to compare two population means, we may wish to compare two population proportions. Is the proportion of literate adults in Punjab higher than in Sindh? Does a new advertisement increase the proportion of consumers who recognize the brand, compared with the old one? Is the proportion of women in the workforce higher in urban areas than in rural areas? In each case the parameter of interest is the difference $(\pi_1 - \pi_2)$, and we wish to construct a confidence interval for this difference.

Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions from two independent random samples of sizes n_1 and n_2. The point estimate of $(\pi_1 - \pi_2)$ is the difference of sample proportions $(\hat{p}_1 - \hat{p}_2)$. It can be shown that the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ has mean $(\pi_1 - \pi_2)$ and standard error:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \qquad (3.12)$$

and is approximately normal when both samples are sufficiently large (the conventional rule is that $n_i\pi_i \geq 5$ and $n_i(1-\pi_i) \geq 5$ for $i = 1, 2$). When the population proportions are unknown — as is almost always the case — they are replaced by the sample proportions, and the $(1-\alpha) \times 100\%$ confidence interval becomes:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \qquad (3.13)$$

Example 3.9 (CI for Difference Between Two Proportions)

A consumer survey was conducted in two cities to compare brand awareness for a particular soft drink. In Lahore, a random sample of $n_1 = 500$ consumers showed that 320 recognized the brand. In Karachi, a random sample of $n_2 = 600$ consumers showed that 348 recognized the brand. Construct a 95% confidence interval for the difference $(\pi_1 - \pi_2)$ in brand recognition between the two cities.

Solution. The two sample proportions are:

$$\hat{p}_1 = \frac{320}{500} = 0.640 \quad , \quad \hat{p}_2 = \frac{348}{600} = 0.580.$$

Both samples easily satisfy the conditions for the normal approximation. The standard error of the difference is:

$$SE = \sqrt{\frac{(0.640)(0.360)}{500} + \frac{(0.580)(0.420)}{600}}$$

$$= \sqrt{0.000461 + 0.000406} = \sqrt{0.000867} = 0.02944.$$

For a 95% confidence interval, $z_{0.025} = 1.96$, and the margin of error is $1.96 \times 0.02944 = 0.0577$. Applying equation (3.13):

$$CI = (0.640 - 0.580) \pm 0.0577 = 0.060 \pm 0.0577 = (0.0023\,;\, 0.1177).$$

We can state with 95% confidence that brand recognition for this soft drink is between 0.2% and 11.8% higher in Lahore than in Karachi. Notice that, although the entire interval lies above zero (suggesting that Lahore really does have higher brand recognition), the lower end of the interval is

very close to zero. The evidence for a real difference is therefore present but not overwhelming. A larger sample would be needed to pin down the difference more precisely.

3.11 CONFIDENCE INTERVAL FOR A POPULATION VARIANCE: THE CHI-SQUARE DISTRIBUTION

In some applications the parameter of interest is not a mean or a proportion but the variance (or, equivalently, the standard deviation). For example, in industrial quality control, the variance of a manufactured component's dimensions is often more important than the mean, because excessive variability is what causes products to fail. In financial analysis, the variance of a stock's returns is a measure of its risk. In educational research, the variance of test scores is sometimes used as a measure of how heterogeneous a class is. In all such cases we wish to construct a confidence interval for the population variance σ^2.

3.11.1 The Chi-Square Distribution

The construction of a confidence interval for σ^2 rests on the following result. If a random sample of size n is drawn from a normal population with variance σ^2, then the quantity

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (3.14)$$

follows a chi-square (χ^2) distribution with $(n - 1)$ degrees of freedom. The chi-square distribution is a probability distribution that is defined only for non-negative values, is right-skewed, and has a shape that depends on its degrees of freedom. As the degrees of freedom increase, the chi-square distribution becomes more and more symmetric and approaches a normal distribution. The critical values of the chi-square distribution for various degrees of freedom and various tail probabilities are tabulated in standard chi-square tables, which can be found in any introductory statistics textbook.

3.11.2 Confidence Interval Formula

Because the chi-square distribution is not symmetric, the construction of a confidence interval for σ^2 requires two different critical values: one for the lower tail and a different one for the upper tail. The $(1 - \alpha) \times 100\%$ confidence interval for σ^2 is given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) \qquad (3.15)$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are the values from the chi-square table with $(n - 1)$ degrees of freedom that cut off $\alpha/2$ in the upper and lower tails respectively. To obtain a confidence interval for the standard deviation σ rather than for the variance σ^2, we simply take the square root of both endpoints:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right) \qquad (3.16)$$

Important caveat. Unlike the confidence intervals we have constructed for means and proportions, the chi-square interval for σ^2 is not protected by the Central Limit Theorem. It depends critically on the assumption that the population itself is normal. When this assumption fails, the chi-square interval can be seriously inaccurate, even for large samples. The student should therefore use this interval only when there is good reason to believe the population is approximately normal — for example, after inspecting a histogram of the sample.

Example 3.10 (Confidence Interval for a Population Variance)

A quality-control engineer at a textile mill in Faisalabad takes a random sample of n = 20 cotton threads and measures their breaking strength (in kilograms). The sample variance is found to be $s^2 = 4.0$ kg². Assuming that the breaking strengths of cotton threads are normally distributed, construct a 95% confidence interval for the population variance σ^2 of breaking strength.

Solution. Here n = 20, $s^2 = 4.0$, and df = n − 1 = 19. For a 95% confidence interval, α = 0.05, so we need $\chi^2_{0.025,19}$ and $\chi^2_{0.975,19}$. From the chi-square table:

$\chi^2_{0.025,19} = 32.852$ (the upper-tail critical value)

$\chi^2_{0.975,19} = 8.907$ (the lower-tail critical value)

Applying equation (3.15):

$$CI = \left(\frac{(19)(4.0)}{32.852}, \frac{(19)(4.0)}{8.907}\right) = \left(\frac{76}{32.852}, \frac{76}{8.907}\right) = (2.313, 8.532).$$

The 95% confidence interval for σ^2 is approximately (2.31 ; 8.53) kg². Taking square roots gives the 95% confidence interval for the standard deviation σ:

$$CI(\sigma) = (\sqrt{2.313}, \sqrt{8.532}) = (1.521, 2.921).$$

We can therefore state with 95% confidence that the population standard deviation of breaking strength lies between approximately 1.52 and 2.92 kilograms. Notice that the interval is rather wide, even though the sample size is 20 — this is a typical characteristic of confidence intervals for variances, which require quite large samples to achieve the same precision that comparable intervals for means achieve with much smaller samples.

3.12 DETERMINING THE REQUIRED SAMPLE SIZE

One of the most common practical questions in any sample-survey project is: "how large a sample do I need?" The answer depends on three things: the desired margin of error (how precise the estimate should be), the desired confidence level (how confident we want to be), and an estimate of the variability in the population. Once these three quantities are specified, the formulas in this section give the smallest sample size that will achieve the required precision.

3.12.1 Sample Size for Estimating a Mean

Recall from Section 3.6 that the margin of error for the (1 − α) × 100% confidence interval for a population mean (with σ known) is:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

We want this margin of error to be no larger than some specified value E. Solving for n:

$$n = \frac{(z_{\alpha/2} \cdot \sigma)^2}{E^2} \quad or \quad n = \frac{z^2_{\alpha/2}\,\sigma^2}{E^2} \qquad (3.17)$$

If σ is unknown — as is usually the case — it can be estimated from a small pilot study, from previous research on the same population, or from a rough guess based on the range of plausible values divided by 4

or 6. The result of the formula should always be rounded up to the next whole number, because rounding down would give an interval slightly wider than required.

Example 3.11 (Sample Size for a Mean)

A market researcher in Islamabad wants to estimate the mean monthly expenditure on entertainment by university students with a margin of error of no more than Rs. 150 at 95% confidence. From a pilot study, the standard deviation is estimated to be approximately $\sigma \approx 900$ rupees. What is the smallest sample size that will achieve this precision?

Solution. We have $z_{0.025} = 1.96$, $\sigma = 900$, and $E = 150$. Applying equation (3.17):

$$n = \frac{(1.96 \cdot 900)^2}{150^2} = \frac{(1764)^2}{22500} = \frac{3{,}111{,}696}{22{,}500} = 138.30.$$

Rounding up, the researcher needs a sample of at least n = 139 students. If the desired margin of error were tightened to Rs. 75 (half as much), the required sample size would jump to n = 553 — about four times as many — confirming the square-root law from Unit 2: cutting the margin of error in half costs four times the observations.

3.12.2 Sample Size for Estimating a Proportion

A similar formula applies for estimating a population proportion. The margin of error for the $(1 - \alpha) \times 100\%$ confidence interval for π is:

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Solving for n:

$$n = \frac{z_{\alpha/2}^2 \, \pi(1 - \pi)}{E^2} \qquad (3.18)$$

When π is unknown in advance — as is usually the case — the conservative practice is to set $\pi = 0.5$, because this maximizes the product $\pi(1 - \pi)$ and therefore gives the largest (and safest) required sample size. With $\pi = 0.5$, the formula simplifies to:

$$n = \frac{z_{\alpha/2}^2}{4E^2} \qquad (3.19)$$

Example 3.12 (Sample Size for a Proportion)

A pollster in Karachi wants to estimate the proportion of voters who support a particular candidate, with a margin of error of no more than 3% (that is, 0.03) at 95% confidence. (a) Compute the required sample size assuming the conservative value $\pi = 0.5$. (b) Recompute the required sample size if a previous poll suggests that the true proportion is approximately $\pi \approx 0.40$.

Solution.

(a) Using equation (3.19) with $z_{0.025} = 1.96$ and $E = 0.03$:

$$n = \frac{(1.96)^2}{4 \cdot (0.03)^2} = \frac{3.8416}{4 \cdot 0.0009} = \frac{3.8416}{0.0036} = 1067.11.$$

Rounding up, the pollster needs a sample of at least n = 1,068 voters. This is the famous "1,000-respondent rule" that explains why most national opinion polls aim for sample sizes of around one thousand.

(b) Using equation (3.18) with $\pi = 0.40$:

$$n = \frac{(1.96)^2 \cdot (0.40)(0.60)}{(0.03)^2} = \frac{3.8416 \cdot 0.24}{0.0009} = \frac{0.92198}{0.0009} = 1024.43.$$

Rounding up, n = 1,025 voters. The required sample size is slightly smaller than in part (a), because the variability of the sample proportion is slightly smaller when $\pi = 0.40$ than when $\pi = 0.50$. The savings, however, are quite modest. This is why the conservative choice $\pi = 0.5$ is so often recommended in practice: it gives the safest answer and almost never costs much in terms of extra sample size.

3.13 SUMMARY

This unit has developed the theory and practice of estimation, one of the two main branches of statistical inference. The main points may be summarized as follows.

- A point estimator is a sample statistic used as a single-number guess for an unknown population parameter; the value it produces for a given sample is called the point estimate. Standard examples include the sample mean $\bar{x}$ for the population mean μ, the sample proportion $\hat{p}$ for the population proportion π, and the sample variance s^2 for the population variance σ^2. A point estimate alone gives no indication of its precision and is therefore of limited usefulness.
- Four desirable properties of a good point estimator are unbiasedness ($E(\hat{\theta}) = \theta$), consistency ($\hat{\theta} \to \theta$ as $n \to \infty$), efficiency (smallest possible variance among unbiased estimators), and sufficiency (uses all the information in the sample about θ). Standard estimators such as $\bar{x}$, $\hat{p}$ and s^2 possess all four properties for the populations most commonly encountered in social-science research.
- The three main methods of constructing point estimators are the method of moments (set sample moments equal to population moments), the method of maximum likelihood (find the parameter value that makes the observed sample most likely), and the method of least squares (minimize the sum of squared deviations from the model). All three methods give the sample mean as the estimator of the population mean.
- A confidence interval is a range of plausible values for an unknown parameter, computed from a sample, together with a stated confidence level (such as 90%, 95% or 99%). The general form of a confidence interval is "estimate ± critical value × standard error". The interval is interpreted in terms of the long-run reliability of the procedure, not as a probability statement about any single interval.
- For a single population mean, the $(1 - \alpha) \times 100\%$ confidence interval is $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ when σ is known, or $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$ when σ is unknown. The t-distribution is used in the second case to allow for the additional uncertainty introduced by estimating σ.

- For the difference between two means, the corresponding interval uses either the pooled-variance t formula (when the two populations are assumed to have equal variances) or Welch's formula (when the variances are clearly unequal).
- For a single proportion, the $(1 - \alpha) \times 100\%$ confidence interval is $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{[\hat{p}(1-\hat{p})/n]}$, valid when both $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$. For the difference between two proportions, the analogous formula uses the standard error formula in equation (3.13).
- For a single population variance, the confidence interval is constructed using the chi-square distribution and is valid only under the assumption of an approximately normal population. The interval is asymmetric because the chi-square distribution is itself asymmetric.
- To determine the sample size needed to estimate a mean with margin of error E at confidence level $(1 - \alpha)$, use $n = (z_{\alpha/2} .\sigma / E)^2$. For a proportion, use $n = z^2_{\alpha/2}\, \pi(1 - \pi) / E^2$, or, conservatively, $n = z^2_{\alpha/2} / (4E^2)$ when π is unknown. The required sample size grows as the square of the desired precision: cutting the margin of error in half costs four times the observations.

In Units 4 and 5 we shall apply exactly the same machinery — the standard normal table, the t-table, the chi-square table, and the corresponding formulas for the standard error — to the closely related problem of testing hypotheses about population parameters. The student who has mastered the material of this unit will find the transition to hypothesis testing very natural.

3.14 SELF-ASSESSMENT EXERCISE

1. Distinguish clearly between an estimator, an estimate and a parameter. Give an example of each in the context of estimating the mean monthly household income in a city.
2. State the four desirable properties of a good point estimator. For each property, give a one-sentence explanation in plain language.
3. What is the difference between a point estimate and an interval estimate? Why is an interval estimate generally more useful in practice?
4. Explain in your own words what is meant by a "95% confidence interval". State both the technically correct interpretation and the looser everyday phrasing, and explain the difference between them.
5. When estimating a population mean, when should the standard normal distribution be used and when should the t-distribution be used? Explain why the t-distribution is needed in the second case.
6. Discuss in detail the four desirable properties of a point estimator (unbiasedness, consistency, efficiency, sufficiency). For each property, give a precise mathematical definition, explain its intuitive meaning, and give an example of an estimator that does or does not possess the property. Show numerically (using a small population if necessary) that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ.
7. Develop the formulas for the $(1 - \alpha) \times 100\%$ confidence intervals for (a) a single population mean when σ is known, (b) a single population mean when σ is unknown, (c) the difference between two population means assuming equal variances, (d) a single population proportion, and (e) the difference between two population proportions. State the assumptions required in each case, and illustrate the use of each formula with at least one numerical example of your own choosing.
8. A random sample of $n = 100$ employees of a multinational company in Karachi gave a mean monthly salary of $\bar{x} = 78{,}000$ rupees. The population standard deviation is known to be $\sigma = 14{,}000$ rupees.

Construct (a) a 90% and (b) a 99% confidence interval for the mean monthly salary μ of all employees of the company. [Answers: (a) (75,697 ; 80,303); (b) (74,396 ; 81,604)]

9. A random sample of n = 25 first-year MBA students at a business school in Lahore gave a mean GMAT score of $\bar{x} = 580$ with a sample standard deviation of s = 60. Assuming GMAT scores are approximately normally distributed, construct a 95% confidence interval for the mean GMAT score μ of all first-year MBA students at the school. [Answer: (555.2 ; 604.8)]

10. A random sample of n = 64 packages handled by a courier company in Islamabad gave a mean delivery time of $\bar{x} = 36.5$ hours with a sample standard deviation of s = 8.2 hours. Construct a 95% confidence interval for the mean delivery time μ. [Answer: $\approx$ (34.45 ; 38.55) hours]

11. A random sample of 800 voters in Punjab found that 480 supported a particular policy. (a) Compute the sample proportion. (b) Construct a 95% confidence interval for the population proportion π of voters supporting the policy. [Answers: (a) $\hat{p} = 0.60$; (b) (0.5661 ; 0.6339)]

12. In a study comparing two teaching methods, $n_1 = 30$ students taught by Method A had a mean test score of $\bar{x}_1 = 78$ with $s_1 = 8$, while $n_2 = 35$ students taught by Method B had a mean of $\bar{x}_2 = 72$ with $s_2 = 9$. Assuming equal variances, construct a 95% confidence interval for the difference $(\mu_1 - \mu_2)$. [Answer: $\approx$ (1.78 ; 10.22)]

13. In a survey of brand recognition, 165 out of 250 consumers in Lahore (sample 1) and 138 out of 220 consumers in Karachi (sample 2) said they recognized a particular brand. Construct a 95% confidence interval for the difference $(\pi_1 - \pi_2)$ in brand recognition between the two cities. [Answer: $\approx (-0.0306 ; 0.1606)$]

14. A quality-control inspector at a tile factory in Gujranwala measures the thickness of a random sample of n = 25 tiles and finds a sample variance of $s^2 = 0.36$ mm². Assuming tile thickness is normally distributed, construct a 95% confidence interval for the population variance σ^2. Then use the result to construct a 95% confidence interval for the standard deviation σ. [Answers: $\sigma^2 \in (0.220 ; 0.696)$; $\sigma \in (0.469 ; 0.834)$ mm]

15. A researcher wants to estimate the mean monthly expenditure on books and stationery by undergraduate students with a margin of error of no more than Rs. 100 at 95% confidence. From past studies, the population standard deviation is estimated to be $\sigma \approx 800$ rupees. What is the smallest sample size that will achieve this precision? [Answer: n = 246]

16. A pollster wants to estimate the proportion of households in Pakistan that have access to clean drinking water, with a margin of error of no more than 4% at 99% confidence. Using the conservative value $\pi = 0.5$, find the required sample size. [Answer: n = 1,037]

17. A random sample of n = 36 school children in a rural district gave a mean weight of $\bar{x} = 28.4$ kg with a sample standard deviation of s = 4.6 kg. Construct a 90% confidence interval for the mean weight μ of school children in the district. [Answer: $\approx$ (27.10 ; 29.70) kg]

UNIT 4: TESTING HYPOTHESES: ONE SAMPLE TESTS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Explain in plain language what a statistical hypothesis is, formulate the null and alternative hypotheses for a given research question, and decide whether the test should be one-tailed or two-tailed.
2. Distinguish clearly between Type I and Type II errors, and explain the meaning and the role of the level of significance α and of the critical region.
3. Apply the standard six-step procedure to carry out a one-sample test of hypothesis for a population mean (with σ known and with σ unknown) and for a population proportion (large samples).
4. Compute and interpret a p-value, and explain how the p-value approach gives the same conclusions as the critical-value approach but is generally more informative.
5. Compute the power of a hypothesis test for a specified value of the alternative parameter, and explain why the power of a test increases as the sample size increases or as the level of significance is relaxed.

4.1 INTRODUCTION

In Unit 3 we developed the theory of estimation, the first of the two great branches of statistical inference. We learned how to use a sample to construct a single-number guess (a point estimate) and a range of plausible values (a confidence interval) for an unknown population parameter. The present unit takes up the second great branch of statistical inference: hypothesis testing. The two branches are closely related and rest on exactly the same theoretical foundation — the sampling distribution that we developed in Unit 2 — but they answer different practical questions. Estimation answers the question "what is the value of the parameter?", whereas hypothesis testing answers the question "is the value of the parameter equal to (or greater than, or less than) some specified number?"

Hypothesis testing is the most widely used inferential technique in the social sciences, and it lies at the heart of nearly every quantitative research paper published in the field. Whenever a researcher writes "the difference was statistically significant at the 5% level", or "the result was significant with p = 0.02", or "we rejected the null hypothesis in favour of the alternative", he or she is reporting the outcome of a hypothesis test. The vocabulary may at first seem strange and the underlying logic may seem indirect, but with a little patience the student will discover that hypothesis testing rests on the same simple idea as a confidence interval: the sampling distribution of the test statistic tells us how likely a particular sample result would be if the assumed value of the parameter were true.

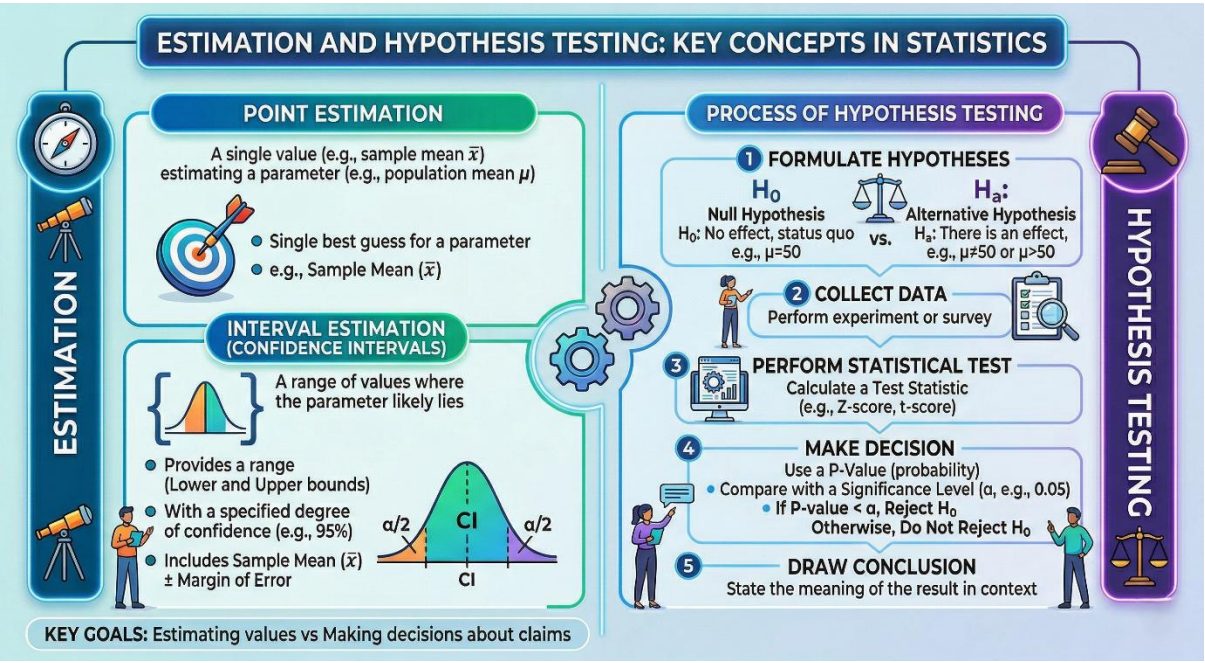


Figure 4.1 Difference between Estimation and Testing

In this unit we restrict our attention to one-sample tests — that is, tests in which a single sample has been drawn from a single population, and the researcher wishes to test a claim about a single population parameter. Two-sample tests, in which two samples are compared with one another, will be the subject of Unit 5. The unit is organized as follows. Sections 4.2 to 4.7 lay down the basic ideas: the role of hypothesis testing in research, how to formulate the hypotheses, the two kinds of error that can be made, the level of significance, the critical region, the distinction between one-tailed and two-tailed tests, and the standard six-step procedure that is used in nearly every test we shall encounter. Sections 4.8 to 4.10 then apply this procedure to the three most important one-sample tests: the test for a population mean with σ known (the Z-test), the test for a

population mean with σ unknown (the t-test), and the test for a population proportion. Section 4.11 introduces the modern p-value approach, which is now standard in computer-based statistical analysis. Section 4.12 develops the concept of the power of a test. The unit closes with a summary and a self-assessment exercise.

4.2 THE ROLE OF STATISTICAL HYPOTHESIS TESTING

A statistical hypothesis is a claim or assertion about the value of one or more parameters of a population. In any practical research project, the hypotheses are not invented by the statistician but arise from the substantive question that the researcher is trying to answer. A few examples will make this clear.

1. A pharmaceutical company claims that a new pain-relief tablet works in less than 15 minutes on average. The hypothesis being tested is "the mean time-to-relief is 15 minutes" against the alternative "the mean time-to-relief is less than 15 minutes".
2. An education researcher believes that a new teaching method produces higher mean test scores than the standard method. The hypothesis being tested is "the new method gives the same mean as the standard method" against the alternative "the new method gives a higher mean than the standard method".
3. A consumer-protection agency wishes to verify that the average net weight of packages sold by a food manufacturer is at least 500 grams, as stated on the label. The hypothesis being tested is "the mean net weight is 500 grams" against the alternative "the mean net weight is less than 500 grams".
4. A pollster claims that more than 60% of voters in a constituency will vote for a particular candidate. The hypothesis being tested is "the population proportion is 60%" against the alternative "the population proportion is greater than 60%".

In each of these cases the researcher has a specific numerical claim in mind, and the role of hypothesis testing is to decide — on the basis of sample evidence — whether the claim is reasonable or not. A statistical test is therefore a formal procedure for using sample data to assess the plausibility of a claim about the population. The procedure is designed in such a way that the long-run probabilities of various kinds of errors can be controlled, so that the researcher knows precisely how much faith to place in the conclusion.

It is important to be clear from the outset about what hypothesis testing can and cannot do. A hypothesis test cannot prove a claim to be true beyond all doubt, because the conclusion is always based on a sample and the sample may, by bad luck, happen to be unrepresentative of the population. What a hypothesis test can do is to give a quantitative measure of how strongly the sample evidence contradicts (or fails to contradict) the claim being tested, and to control the long-run rate at which the test will lead to wrong conclusions. A statistically significant result is not a logical proof; it is a statement that, if the null hypothesis were really true, the observed sample result would be very unlikely.

4.3 FORMULATING THE NULL AND ALTERNATIVE HYPOTHESES

The first step in any hypothesis test is to write down two competing statements about the parameter of interest, called the null hypothesis and the alternative hypothesis. These two hypotheses must be mutually exclusive (they cannot both be true at the same time) and exhaustive (between them they cover all possibilities). The notation used is universal: the null hypothesis is denoted by H_0 (read as "H-naught" or "H-zero") and the alternative hypothesis by H_1 or H_a.

4.3.1 The Null Hypothesis

The null hypothesis H_0 is the statement that is being tested. It typically expresses a state of "no difference", "no effect" or "no change", and it is the statement that the researcher will retain unless the sample evidence forces him or her to abandon it. The null hypothesis always contains an equality sign in its mathematical statement — for example, $H_0: \mu = 50$ or $H_0: \pi = 0.5$. (Some textbooks use $\leq$ or $\geq$ in the null hypothesis when the alternative is one-tailed, but the equality form is the simplest and is the one we shall adopt throughout this unit.) The role of the null hypothesis is rather similar to that of the "presumption of innocence" in a criminal trial: the accused is presumed innocent until proven guilty, and the null hypothesis is presumed true until the sample evidence proves it false beyond reasonable doubt.

4.3.2 The Alternative Hypothesis

The alternative hypothesis H_1 is the statement that the researcher really wants to establish. It is the rival claim that will be accepted if the null hypothesis is rejected. The alternative hypothesis can take one of three forms:

1. Two-sided (or two-tailed): $H_1: \mu \neq \mu_0$. This form is used when the researcher wishes to detect any departure from the value μ_0, in either direction.
2. Right-tailed (or upper-tailed): $H_1: \mu > \mu_0$. This form is used when the researcher wishes to detect only a departure to values greater than μ_0.
3. Left-tailed (or lower-tailed): $H_1: \mu < \mu_0$. This form is used when the researcher wishes to detect only a departure to values less than μ_0.

The choice between these three forms depends entirely on the substantive research question, not on the data themselves. The researcher must decide in advance — before looking at the sample — what kind of departure from H_0 would be of interest. Choosing the form of the alternative after looking at the data is a serious methodological error that invalidates the whole procedure.

Example 4.1 (Formulating Hypotheses)

Translate each of the following research questions into a pair of null and alternative hypotheses about the appropriate population parameter, and identify whether the test is one-tailed or two-tailed.

(a) A nutritionist believes that the average daily calorie intake of adult women in Punjab is exactly 2,000 calories. A random sample is drawn to verify this claim.

Solution. Let μ denote the mean daily calorie intake. The hypotheses are $H_0: \mu = 2{,}000$ versus $H_1: \mu \neq 2{,}000$. The test is two-tailed because the researcher is interested in any departure from 2,000 in either direction.

(b) A car-tyre manufacturer claims that his tyres last on average more than 50,000 km. A consumer-protection agency wishes to test this claim.

Solution. Let μ denote the mean tyre lifetime. The agency wishes to know whether the manufacturer's claim ($\mu > 50{,}000$) is supported. The hypotheses are $H_0: \mu = 50{,}000$ versus $H_1: \mu > 50{,}000$. The test is one-tailed (right-tailed).

(c) A researcher in Karachi believes that fewer than 30% of households have access to a public-sewerage system. A random sample of households is drawn.

Solution. Let π denote the population proportion of households with public-sewerage access. The hypotheses are H_0: $\pi = 0.30$ versus H_1: $\pi < 0.30$. The test is one-tailed (left-tailed).

In each case the form of the alternative is dictated by the substantive question, not by the data. A student should be able to write down the hypotheses for a given research problem before any sample is even drawn.

4.4 TYPE I AND TYPE II ERRORS

Whenever we use a sample to make a decision about a population, there is a chance that our decision will be wrong, simply because the sample contains less information than the population. The key insight of hypothesis testing — first articulated formally by Jerzy Neyman and Egon Pearson in the 1930s — is that there are exactly two ways in which a hypothesis test can lead to a wrong conclusion. These are called Type I and Type II errors.

4.4.1 Definitions

A Type I error is the error of rejecting the null hypothesis H_0 when in fact H_0 is true. In other words, a Type I error is a "false alarm": the test decides that something interesting is going on when in fact it isn't. A Type II error is the error of failing to reject the null hypothesis H_0 when in fact H_0 is false. In other words, a Type II error is a "miss": the test fails to detect a real effect that does exist.

The two errors and the two correct decisions can be summarized in the following 2 × 2 table:

	H_0 is true	**H_0 is false**
Reject H_0	Type I error (probability α)	Correct decision (probability $1 - \beta$)
Fail to reject H_0	Correct decision (probability$1 - \alpha$)	Type II error (probability β)

The probability of a Type I error is universally denoted by α (the Greek letter alpha) and is also called the level of significance of the test. The probability of a Type II error is denoted by β (the Greek letter beta). The complementary probability $(1 - \beta)$ is called the power of the test, and it represents the probability of correctly rejecting H_0 when H_0 is in fact false. We shall return to the power of a test in Section 4.12.

4.4.2 The Trade-Off Between α and β

A natural question is: why not simply make both α and β as small as possible? The answer is that, for any fixed sample size, decreasing α automatically increases β, and vice versa. The two error probabilities are linked, and there is an unavoidable trade-off between them. The only way to reduce both error probabilities at the same time is to increase the sample size.

The usual practice in statistical testing is to specify the level of significance α in advance — often setting it at 0.05 or 0.01 — and then to design the test so as to minimize β subject to that constraint. The choice of α reflects the cost of a Type I error: if a false alarm would have very serious consequences (for example, in the testing of a new drug), a small value such as $\alpha = 0.01$ or even 0.001 is appropriate. If the cost of a false alarm is more modest, the conventional value $\alpha = 0.05$ is usually adequate. There is no scientific reason for these particular numbers; they are conventions established by long usage.

4.5 LEVEL OF SIGNIFICANCE AND CRITICAL REGION

The level of significance α is the maximum probability of a Type I error that the researcher is willing to tolerate. Once α has been chosen, the next step in the hypothesis test is to use α to determine the critical region (also called the rejection region), which is the set of values of the test statistic for which the null hypothesis will be rejected. The complementary set is called the acceptance region (or, more accurately, the non-rejection region).

To be concrete, consider a Z-test of $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ at level $\alpha = 0.05$. The test statistic

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

follows the standard normal distribution $N(0, 1)$ under H_0. For $\alpha = 0.05$, the critical region of a two-tailed Z-test consists of all values of Z that are either less than -1.96 or greater than $+1.96$. The total area of the standard normal curve in the two tails beyond ± 1.96 is exactly 0.05. If the value of Z computed from our sample falls in this critical region, we reject H_0; if it falls in the central region between -1.96 and $+1.96$, we fail to reject H_0.

For a one-tailed test the critical region is concentrated entirely in one tail of the distribution. For a right-tailed test ($H_1: \mu > \mu_0$) at $\alpha = 0.05$, the critical region is $z > 1.645$, with the entire 5% area in the right tail. For a left-tailed test ($H_1: \mu < \mu_0$) at $\alpha = 0.05$, the critical region is $z < -1.645$, with the entire 5% area in the left tail. The most commonly used critical values are summarized below:

Significance Level α	Two-Tailed	Right-Tailed	Left-Tailed
0.10	± 1.645	$+1.28$	-1.28
0.05	± 1.96	$+1.645$	-1.645
0.01	± 2.58	$+2.33$	-2.33
0.001	± 3.29	$+3.09$	-3.09

4.6 ONE-TAILED AND TWO-TAILED TESTS

We have already met the distinction between one-tailed and two-tailed tests in connection with the form of the alternative hypothesis. We now elaborate on the practical and conceptual differences between the two kinds of test.

A two-tailed test is appropriate when the researcher is interested in detecting departures from H_0 in either direction. The classic example is the verification of a manufacturing process: the engineer wants to detect any drift in the mean diameter of bolts, whether the bolts have become too thick or too thin, because both kinds of departure are equally undesirable. The critical region of a two-tailed test is split equally between the two tails of the sampling distribution, with $\alpha/2$ in each tail.

A one-tailed test is appropriate when only one direction of departure is of interest. There are two common situations in which one-tailed tests are used. The first is when the researcher has a strong prior reason to expect departures in only one direction — for example, when testing whether a new fertilizer increases (rather than just changes) crop yield, or whether a new advertising campaign raises (rather than just affects) sales. The second is when the consequences of departures in one direction are negligible and only departures in the other direction matter — for example, when verifying that the contents of a package are at least the labelled amount; nobody complains when there is more, only when there is less.

A one-tailed test is more powerful than the corresponding two-tailed test at the same level of significance, because the entire α is concentrated in one tail rather than being split between the two. The price of this extra power is that a one-tailed test cannot detect departures in the wrong direction at all. A researcher who runs a right-tailed test at $\alpha = 0.05$ and obtains a sample mean far below the hypothesized value cannot reject H_0, even if the departure would have been overwhelming evidence in a two-tailed test. For this reason it is sometimes said, half-seriously, that a one-tailed test should be used only when the researcher would not even bother to run an experiment to detect the other direction. A safer general rule for beginning students is: when in doubt, use a two-tailed test.

4.7 GENERAL PROCEDURE FOR A HYPOTHESIS TEST

Almost every hypothesis test we shall encounter in this course follows the same six-step procedure. The student is strongly advised to commit these six steps to memory and to use them as a checklist for every test, no matter what the underlying parameter or distribution may be. The discipline of going through every step in order is a powerful safeguard against the common mistakes that beginners make.

1. State the null and alternative hypotheses clearly, in symbolic form, and identify whether the test is one-tailed or two-tailed.
2. Choose the level of significance α. The conventional choice is $\alpha = 0.05$ unless there is a specific reason to use a different value.
3. Identify the appropriate test statistic for the situation, and write down its sampling distribution under the assumption that H_0 is true. (The choice of test statistic depends on the parameter being tested and on what is known about the population — for example, whether σ is known or unknown.)
4. Determine the critical region for the chosen α and the chosen form of the alternative. This involves looking up the appropriate critical value from the standard normal table, the t-table, or some other table.
5. Compute the value of the test statistic from the sample data, substituting the value of the parameter specified by H_0.
6. Compare the computed value of the test statistic with the critical region, and reach a decision: reject H_0 if the computed value falls in the critical region, otherwise fail to reject H_0. State the conclusion in the language of the original research problem, not just in symbolic form.

When the same procedure is followed for many different tests, the student soon develops a feel for the underlying logic, and the formulas for different tests stop seeming arbitrary. The next three sections apply this six-step procedure to the three most important one-sample tests.

4.8 TESTING THE MEAN OF A NORMAL POPULATION (σ KNOWN): THE Z-TEST

We now apply the six-step procedure to the simplest one-sample test: testing a hypothesis about a population mean μ when the population standard deviation σ is known. This case is somewhat artificial in practice, because if we know σ we very often also know μ. Nevertheless, the Z-test is the natural starting point: the formulas are clean, the underlying logic is easy to follow, and once the Z-test is mastered the more realistic t-test of the next section becomes a straightforward generalization.

4.8.1 The Test Statistic

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a normal population with unknown mean μ and known standard deviation σ. (If the population is not normal but the sample size is large enough — say, $n \geq 30$ — the Central Limit Theorem allows us to use the same procedure as an approximation.) The null hypothesis specifies a particular value μ_0 for the mean: $H_0: \mu = \mu_0$. Under this null hypothesis, the sampling distribution of $\bar{x}$ is normal with mean μ_0 and standard error $\sigma/\sqrt{n}$. The standardized test statistic is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad (4.1)$$

and follows the standard normal distribution N(0, 1) under H_0. The critical region depends on the form of the alternative hypothesis, as summarized in the following table:

Alternative Hypothesis	Form of the Critical Region	Critical Region for α = 0.05
$H_1: \mu \neq \mu_0$ (two-tailed)	$\lvert Z \rvert > z_{\alpha/2}$	$\lvert Z \rvert > 1.96$
$H_1: \mu > \mu_0$ (right-tailed)	$Z > z_\alpha$	$Z > 1.645$
$H_1: \mu < \mu_0$ (left-tailed)	$Z < -z_\alpha$	$Z < -1.645$

Example 4.2 (Z-test, Two-tailed, σ Known)

A factory in Faisalabad produces steel rods that are supposed to have a mean length of 100 cm. The manufacturing process is known to produce rods whose lengths follow a normal distribution with population standard deviation σ = 2 cm. In a routine quality-control check, a random sample of n = 25 rods is drawn from the day's production and gives a sample mean of $\bar{x}$ = 100.8 cm. Test, at the 5% level of significance, whether the production process is operating at the correct mean length.

Solution. We follow the six-step procedure.

Step 1. Hypotheses. Let μ denote the true mean length of the rods produced today. We wish to detect any departure from the target value of 100 cm in either direction:

$$H_0: \mu = 100 \quad versus \quad H_1: \mu \neq 100 \quad (two-tailed\ test).$$

Step 2. Level of significance: α = 0.05.

Step 3. Test statistic. Because σ is known and the population is normal, we use the Z-test (equation 4.1):

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)\ under\ H_0.$$

Step 4. Critical region. For a two-tailed test at α = 0.05, the critical values are $\pm z_{0.025} = \pm 1.96$. We reject H_0 if $|Z| > 1.96$.

Step 5. Computation:

$$Z = \frac{100.8 - 100}{2/\sqrt{25}} = \frac{0.8}{2/5} = \frac{0.8}{0.4} = 2.0.$$

Step 6. Decision and conclusion. Since $|Z| = 2.0 > 1.96$, the computed value falls in the critical region. We reject H_0 at the 5% level of significance. There is sufficient evidence to conclude that the true mean length of the rods produced today differs from the target value of 100 cm. The quality-

control team should investigate the production process and bring the mean length back into specification.

Example 4.3 (Z-test, Right-tailed, σ Known)

A telecommunications company in Lahore claims that the average data download speed on its 4G network is at least 25 Mbps. A consumer-protection group, sceptical of this claim, draws a random sample of n = 36 download tests at different times and locations and obtains a sample mean of $\bar{x}$ = 23.5 Mbps. The population standard deviation of download speeds is known from earlier studies to be σ = 6 Mbps. Test, at the 5% level of significance, whether the company's claim is justified.

Solution.

Step 1. Hypotheses. The claim is that the mean is at least 25, so the consumer group will look for evidence that the mean is in fact less than 25. We therefore set:

$$H_0: \mu = 25 \quad versus \quad H_1: \mu < 25 \quad (left - tailed\ test).$$

Step 2. α = 0.05.

Step 3. Test statistic: the Z-test (4.1).

Step 4. Critical region. For a left-tailed test at α = 0.05, the critical value is $-z_{0.05} = -1.645$. We reject H_0 if $Z < -1.645$.

Step 5. Computation:

$$Z = \frac{23.5 - 25}{6/\sqrt{36}} = \frac{-1.5}{6/6} = \frac{-1.5}{1.0} = -1.50.$$

Step 6. Decision and conclusion. Since $Z = -1.50$ is greater than -1.645, the computed value does not fall in the critical region. We fail to reject H_0 at the 5% level of significance. There is not sufficient evidence to contradict the company's claim that the mean download speed is at least 25 Mbps. Notice that this does not prove that the claim is true; it only means that the present sample does not give us enough evidence to disprove it. With a larger sample, the same sample mean of 23.5 might well lead to a rejection of H_0.

4.9 TESTING THE MEAN OF A NORMAL POPULATION (σ UNKNOWN): THE t-TEST

In nearly every real research situation, the population standard deviation σ is unknown and must be estimated from the same sample as the mean. We met this situation in Section 3.7 in connection with confidence intervals, and we found that we had to switch from the standard normal distribution to the Student's t-distribution to allow for the additional uncertainty introduced by estimating σ. The same switch is needed in hypothesis testing. The procedure is identical to the Z-test of Section 4.8, except that:

1. The known population standard deviation σ is replaced by the sample standard deviation s.
2. The standard normal distribution $N(0, 1)$ is replaced by the Student's t-distribution with $(n - 1)$ degrees of freedom.
3. The critical values are read from the t-table rather than from the standard normal table.

4.9.1 The Test Statistic

The test statistic for testing H_0: $\mu = \mu_0$ when σ is unknown is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad (4.2)$$

which follows Student's t-distribution with df = n − 1 degrees of freedom under H_0. The form of the critical region is the same as for the Z-test, with t replacing z and the t-critical values replacing the corresponding z-values:

Alternative Hypothesis	Form of Critical Region
$H_1: \mu \neq \mu_0$ (two-tailed)	$\lvert t \rvert > t_{\alpha/2,\, n-1}$
$H_1: \mu > \mu_0$ (right-tailed)	$t > t_{\alpha,\, n-1}$
$H_1: \mu < \mu_0$ (left-tailed)	$t < -t_{\alpha,\, n-1}$

Example 4.4 (t-Test, Two-tailed, Small Sample)

A pediatrician at a hospital in Islamabad wishes to test whether the mean haemoglobin level of children aged 5 to 7 years in a particular community differs from the standard value of 12 g/dL. A random sample of n = 16 children from the community gives a sample mean of $\bar{x}$ = 11.4 g/dL with a sample standard deviation of s = 1.2 g/dL. Assuming haemoglobin levels are approximately normally distributed, test the pediatrician's claim at the 5% level of significance.

Solution.

Step 1. Hypotheses. We are interested in any difference from the standard value, in either direction:

$$H_0: \mu = 12 \quad versus \quad H_1: \mu \neq 12 \quad (two-tailed\ test).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: the t-test (4.2), with df = n − 1 = 15.

Step 4. Critical region. For a two-tailed test with df = 15 at $\alpha = 0.05$, the critical values from the t-table are $\pm t_{0.025,\, 15} = \pm 2.131$. We reject H_0 if $|t| > 2.131$.

Step 5. Computation:

$$t = \frac{11.4 - 12}{1.2/\sqrt{16}} = \frac{-0.6}{1.2/4} = \frac{-0.6}{0.3} = -2.0.$$

Step 6. Decision and conclusion. Since |t| = 2.0 is less than 2.131, the computed value does not fall in the critical region. We fail to reject H_0 at the 5% level of significance. There is not sufficient evidence to conclude that the mean haemoglobin level of children in the community differs from the standard value of 12 g/dL. The result is, however, "borderline": the test statistic is very close to the critical value, and a slightly larger sample (or a slightly larger departure from 12) would have changed the conclusion. The pediatrician would be wise to collect more data before drawing any firm policy conclusion.

Example 4.5 (t-Test, Right-tailed)

The principal of a college in Multan claims that the mean monthly income of his college's graduates, five years after graduation, exceeds Rs. 80,000. A random sample of n = 25 graduates gives a sample mean of $\bar{x}$ = 84,500 rupees with a sample standard deviation of s = 12,000 rupees. Test, at the 1% level of significance, whether the principal's claim is supported by the data.

Solution.

Step 1. Hypotheses. The principal claims that μ exceeds 80,000, so:

$$H_0: \mu = 80{,}000 \quad versus \quad H_1: \mu > 80{,}000 \quad (right-tailed\ test).$$

Step 2. α = 0.01.

Step 3. Test statistic: the t-test, with df = 24.

Step 4. Critical region. For a right-tailed test at α = 0.01 with df = 24, the critical value is $t_{0.01,\,24}$ = 2.492. We reject H_0 if t > 2.492.

Step 5. Computation:

$$t = \frac{84500 - 80000}{12000/\sqrt{25}} = \frac{4500}{12000/5} = \frac{4500}{2400} = 1.875.$$

Step 6. Decision and conclusion. Since t = 1.875 is less than 2.492, the computed value does not fall in the critical region. We fail to reject H_0 at the 1% level of significance. There is not sufficient evidence at the 1% level to conclude that the mean monthly income of the college's graduates exceeds Rs. 80,000. Notice, however, that the same sample would have led to rejection at the 5% level (where the critical value is $t_{0.05,\,24}$= 1.711, which is less than 1.875). This illustrates how sensitive the conclusion of a hypothesis test can be to the choice of α — a point we shall return to in connection with the p-value approach in Section 4.11.

Example 4.5(b) (t-Test, Left-tailed)

A bottling plant in Hyderabad fills soft-drink bottles to a labelled volume of 250 ml. A consumer-protection inspector suspects that the bottles are being underfilled. He draws a random sample of n = 20 bottles and finds a mean volume of $\bar{x}$ = 247.6 ml with a sample standard deviation of s = 4.5 ml. Assuming volumes are normally distributed, test at α = 0.05 whether the mean fill volume is less than the labelled 250 ml.

Solution.

Step 1. The inspector wishes to detect under-filling, so the alternative is one-sided in the negative direction:

$$H_0: \mu = 250 \quad versus \quad H_1: \mu < 250 \quad (left-tailed\ test).$$

Step 2. α = 0.05.

Step 3. Test statistic: the t-test (4.2), with df = 19.

Step 4. Critical region. For a left-tailed test at $\alpha = 0.05$ with df = 19, the critical value is $-t_{0.05,\,19} = -1.729$. We reject H_0 if $t < -1.729$.

Step 5. Computation:

$$t = \frac{247.6 - 250}{4.5/\sqrt{20}} = \frac{-2.4}{4.5/4.472} = \frac{-2.4}{1.0062} = -2.385.$$

Step 6. Decision and conclusion. Since $t = -2.385$ is less than -1.729, the test statistic falls in the critical region. We reject H_0 at the 5% level of significance. There is sufficient evidence to conclude that the mean fill volume is less than the labelled 250 ml. The inspector should report his finding to the bottling plant and require corrective action. Note that the test statistic is also less than $-2.539 = -t_{0.01,\,19}$, so the result is in fact significant even at the 1% level.

4.9.2 The Connection Between Confidence Intervals and Hypothesis Tests

There is a beautiful and very useful connection between confidence intervals (which we developed in Unit 3) and two-tailed hypothesis tests (which we have developed in this unit). The connection is the following: a two-tailed test of $H_0: \mu = \mu_0$ at level α leads to rejection of H_0 if and only if the value μ_0 does not lie inside the $(1 - \alpha) \times 100\%$ confidence interval for μ. Equivalently: a value μ_0 that lies inside the 95% confidence interval cannot be rejected as a hypothesis at the 5% level, and a value that lies outside the interval can be rejected.

This duality means that, in principle, we never need to perform both a hypothesis test and a confidence interval calculation on the same data — either one alone gives us all the information we need. In practice, however, the two procedures convey the information in slightly different ways: the confidence interval tells us the entire range of plausible values for the parameter, whereas the hypothesis test tells us whether one specific claimed value is plausible. Many empirical research papers therefore report both, and the student should be comfortable moving back and forth between them.

Example 4.5(c) (Hypothesis Test from a Confidence Interval)

A 95% confidence interval for the mean monthly household expenditure on food in a residential colony is computed to be (Rs. 17,200 ; Rs. 19,800). Without doing any further calculations, decide whether each of the following hypotheses about the true population mean μ would be rejected by a two-tailed test at $\alpha = 0.05$:

(a) $H_0: \mu =$ Rs. 18,000; (b) $H_0: \mu =$ Rs. 17,000; (c) $H_0: \mu =$ Rs. 20,000.

Solution. Using the duality between confidence intervals and two-tailed tests at the corresponding level:

1. (a) The hypothesized value 18,000 lies inside the 95% interval (17,200 ; 19,800), so we fail to reject H_0 at $\alpha = 0.05$.
2. (b) The hypothesized value 17,000 lies outside (below) the 95% interval, so we reject H_0 at $\alpha = 0.05$.
3. (c) The hypothesized value 20,000 also lies outside (above) the 95% interval, so we reject H_0 at $\alpha = 0.05$.

Notice how much information the single confidence interval (17,200 ; 19,800) gives us. Without performing any additional arithmetic, we can immediately answer infinitely many two-tailed hypothesis tests at $\alpha = 0.05$: any value of μ inside the interval is "plausible" and any value outside is "implausible".

4.10 TESTING A POPULATION PROPORTION: LARGE SAMPLES

Many of the most important hypothesis tests in social-science research concern proportions rather than means. Does the proportion of literate adults in a district exceed 60%? Has the support for a particular political party changed since the last election? Does a new medication have a cure rate higher than the standard treatment? In each case the parameter of interest is a population proportion π, and the test statistic is constructed in much the same way as for a population mean.

4.10.1 The Test Statistic

Let $\hat{p} = X/n$ denote the sample proportion of "successes" in a sample of size n drawn from a population in which the true proportion of successes is π. We saw in Section 2.15 that, for large n, the sampling distribution of $\hat{p}$ is approximately normal with mean π and standard error $\sqrt{[\pi(1-\pi)/n]}$. The conventional rule of thumb is that the normal approximation is valid when both $n\pi \geq 5$ and $n(1-\pi) \geq 5$; in this section we assume that this condition holds.

To test $H_0: \pi = \pi_0$ against an appropriate alternative, we use the test statistic:

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \qquad (4.3)$$

which follows the standard normal distribution $N(0,1)$ approximately under H_0. Notice that, in computing the standard error in the denominator, we use the hypothesized value π_0 from the null hypothesis, not the sample proportion $\hat{p}$. This is different from the procedure for confidence intervals, where (because there is no hypothesized value to use) we substitute $\hat{p}$. Both choices are reasonable; using π_0 in the test statistic is the more common convention because it makes the null distribution exactly $N(0,1)$.

The form of the critical region depends on the alternative hypothesis in exactly the same way as for the Z-test of a mean:

Alternative Hypothesis	Form of the Critical Region
$H_1: \pi \neq \pi_0$ (two-tailed)	$\lvert Z \rvert > z_{\alpha/2}$
$H_1: \pi > \pi_0$ (right-tailed)	$Z > z_{\alpha}$
$H_1: \pi < \pi_0$ (left-tailed)	$Z < -z_{\alpha}$

Example 4.6 (Test of a Proportion, Right-tailed)

A pollster in Karachi claims that more than 60% of voters in a particular constituency support a certain candidate. To test this claim, an independent research agency draws a random sample of n

= 400 voters from the constituency and finds that 256 of them support the candidate. Test, at the 5% level of significance, whether the pollster's claim is supported by the data.

Solution.

Step 1. Hypotheses. The pollster claims $\pi > 0.60$:

$$H_0: \pi = 0.60 \quad versus \quad H_1: \pi > 0.60 \quad (right-tailed\ test).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: equation (4.3). Both $n\pi_0 = 400(0.60) = 240$ and $n(1-\pi_0) = 400(0.40) = 160$ are well above 5, so the normal approximation is valid.

Step 4. Critical region. For a right-tailed test at $\alpha = 0.05$, the critical value is $z_{0.05} = 1.645$. We reject H_0 if $Z > 1.645$.

Step 5. Computation. The sample proportion is $\hat{p} = 256/400 = 0.64$. The standard error under H_0 is:

$$SE = \sqrt{\frac{(0.60)(0.40)}{400}} = \sqrt{0.0006} = 0.02449.$$

$$Z = \frac{0.64 - 0.60}{0.02449} = \frac{0.04}{0.02449} = 1.633.$$

Step 6. Decision and conclusion. Since $Z = 1.633$ is less than the critical value 1.645 (just barely), we fail to reject H_0 at the 5% level. The sample evidence is not quite strong enough to support the pollster's claim that more than 60% of voters support the candidate. The result is on the very edge — and this is exactly the kind of borderline case where reporting a p-value (Section 4.11) is much more informative than the simple "reject / fail to reject" decision.

Example 4.7 (Test of a Proportion, Two-tailed)

A medical researcher in Lahore claims that exactly 30% of adults in a certain district have hypertension. A health survey of $n = 600$ adults randomly selected from the district finds that 162 of them have hypertension. Test, at the 1% level of significance, whether the researcher's claim is consistent with the survey data.

Solution.

Step 1. Hypotheses. We are interested in any departure from 0.30 in either direction:

$$H_0: \pi = 0.30 \quad versus \quad H_1: \pi \neq 0.30 \quad (two-tailed\ test).$$

Step 2. $\alpha = 0.01$.

Step 3. Test statistic: equation (4.3). Both $n\pi_0 = 180$ and $n(1-\pi_0) = 420$ are well above 5.

Step 4. Critical region. For a two-tailed test at $\alpha = 0.01$, the critical values are $\pm z_{0.005} = \pm 2.576$. We reject H_0 if $|Z| > 2.576$.

Step 5. Computation. The sample proportion is $\hat{p} = 162/600 = 0.27$. The standard error is:

$$SE = \sqrt{\frac{(0.30)(0.70)}{600}} = \sqrt{0.00035} = 0.01871.$$

$$Z = \frac{0.27 - 0.30}{0.01871} = \frac{-0.03}{0.01871} = -1.604.$$

Step 6. Decision and conclusion. Since |Z| = 1.604 is less than 2.576, the computed value does not fall in the critical region. We fail to reject H_0 at the 1% level of significance. There is not sufficient evidence at the 1% level to contradict the researcher's claim that the prevalence of hypertension is 30% in this district. Note: at the more relaxed 10% level (with critical values ±1.645), the result is also not significant — barely. This case shows how a single sample can lead to different conclusions at different levels of significance, reinforcing the importance of choosing α before looking at the data.

Example 4.7(b) (Test of a Proportion, Left-tailed)

A new vaccination program launched in a district of Khyber Pakhtunkhwa aims to bring the proportion of unvaccinated children below 15%. After the program has been running for one year, a random sample of n = 800 children is drawn from the district and 96 are found to be unvaccinated. Test, at α = 0.05, whether the proportion of unvaccinated children is now less than 15%.

Solution.

Step 1. Hypotheses:

$$H_0: \pi = 0.15 \quad versus \quad H_1: \pi < 0.15 \quad (left - tailed\ test).$$

Step 2. α = 0.05.

Step 3. Test statistic: equation (4.3). Both $n\pi_0 = 120$ and $n(1 - \pi_0) = 680$ are well above 5.

Step 4. Critical region. For a left-tailed test at α = 0.05, the critical value is $-z_{0.05} = -1.645$. We reject H_0 if $Z < -1.645$.

Step 5. Computation. The sample proportion is $\hat{p} = 96/800 = 0.12$. The standard error under H_0 is:

$$SE = \sqrt{\frac{(0.15)(0.85)}{800}} = \sqrt{0.00015938} = 0.01262.$$

$$Z = \frac{0.12 - 0.15}{0.01262} = \frac{-0.03}{0.01262} = -2.378.$$

Step 6. Decision and conclusion. Since Z = −2.378 is less than −1.645, the test statistic falls in the critical region. We reject H_0 at the 5% level of significance. There is sufficient evidence to conclude that the proportion of unvaccinated children in the district is now less than 15%. The vaccination program appears to be having its intended effect, and the public-health authorities can be reasonably confident that the target has been met.

4.11 THE p-VALUE APPROACH

The classical hypothesis-testing procedure described in the previous sections gives a simple yes-or-no answer: either the null hypothesis is rejected or it is not. This dichotomy has the virtue of being clean and easy to communicate, but it has one important drawback: it gives no indication of just how strongly the data contradict (or support) the null hypothesis. A test result that just barely falls into the critical region is treated in exactly the same way as one that falls deep inside it; both lead to the same simple decision "reject H_0". This is unsatisfying, because intuitively the second result is much stronger evidence against H_0 than the first. The p-value approach, which has become standard in computer-based statistical analysis, repairs this defect by reporting a continuous measure of the strength of the evidence rather than a binary decision.

4.11.1 Definition of the p-Value

The p-value of an observed test statistic is the probability — computed under the assumption that H_0 is true — of obtaining a test statistic at least as extreme as the one actually observed, in the direction of the alternative hypothesis. In symbols, for a right-tailed test the p-value is $P(Z \geq z_obs \mid H_0)$; for a left-tailed test it is $P(Z \leq z_obs \mid H_0)$; and for a two-tailed test it is $2 \times P(Z \geq |z_obs| \mid H_0)$. The smaller the p-value, the stronger the evidence against H_0.

A loose but useful interpretation of the p-value is "the probability of observing what we observed, or something more extreme, just by chance, if the null hypothesis is really true". When the p-value is very small (say, less than 0.05), the data are unlikely to have arisen by chance under H_0, and so we conclude that H_0 is probably false. When the p-value is large (say, greater than 0.10), the data are quite consistent with H_0, and so we have no reason to reject it. Conventional thresholds for interpreting p-values are summarized below, although the student should remember that these are guidelines, not iron laws.

p-Value	Strength of Evidence Against H_0
$p > 0.10$	No real evidence against H_0
$0.05 < p \leq 0.10$	Weak evidence against H_0 (suggestive)
$0.01 < p \leq 0.05$	Moderate evidence against H_0 (significant)
$0.001 < p \leq 0.01$	Strong evidence against H_0 (highly significant)
$p \leq 0.001$	Very strong evidence against H_0 (extremely significant)

4.11.2 The Decision Rule with p-Values

The p-value approach replaces the comparison of the test statistic with a critical value by a much simpler comparison of the p-value with the chosen level of significance:

1. If $p \leq \alpha$, reject H_0.
2. If $p > \alpha$, fail to reject H_0.

The two approaches always give exactly the same decision at the same level of significance. The advantage of reporting the p-value instead of just the decision is that any reader of the report can apply his or her own preferred level of significance to the same data without redoing the calculations. A reader who normally uses $\alpha = 0.05$ will reach the conclusion "reject" whenever the reported p-value is less than 0.05; a reader who prefers $\alpha = 0.01$ will reach the conclusion "reject" whenever the reported p-value is less than 0.01. Reporting the p-value lets the reader make this decision; reporting only "reject H_0 at the 5% level" does not.

Example 4.8 (Computing a p-Value)

In Example 4.5, we tested the principal's claim that the mean monthly income of college graduates exceeds Rs. 80,000, and computed the test statistic t = 1.875 with df = 24. (a) Compute (or bracket) the p-value of this test. (b) State the conclusion at α = 0.05 and at α = 0.01.

Solution.

(a) From the t-table with $df = 24$, we have $t_{0.05,24} = 1.711$ and $t_{0.05,24} = 2.064$. The observed test statistic 1.875 falls between these two values, so the right-tail p-value lies between 0.025 and 0.05. (Software gives the more precise value $p \approx 0.0366$, but the bracketing is enough for us to make the decision.)

(b) At $\alpha = 0.05$: since $p \approx 0.037 < 0.05$, we reject H_0. The data give moderate evidence that the mean monthly income exceeds Rs. 80,000. At $\alpha = 0.01$: since $p \approx 0.037 > 0.01$, we fail to reject H_0. The data are not strong enough to support the same conclusion at the more demanding 1% level. Notice how the p-value of 0.037 immediately tells us that the result is significant at any level above 0.037 and not significant at any level below 0.037. A single number conveys far more information than the simple "reject / fail to reject" decision at a single fixed α.

Example 4.8(b) (p-Value for a Two-Tailed Z-Test)

In Example 4.2, we tested whether the mean length of steel rods at a Faisalabad factory differed from 100 cm and computed the test statistic Z = 2.0 in a two-tailed Z-test. Compute the p-value of this test and use it to decide whether to reject H_0 at α = 0.05 and at α = 0.01.

Solution. For a two-tailed test, the p-value is twice the area in one tail beyond the absolute value of the test statistic:

$$p = 2 \times P(Z \geq |2.0|) = 2 \times P(Z \geq 2.0).$$

From the standard normal table, $P(Z \geq 2.0) = 1 - 0.9772 = 0.0228$. Therefore:

$$p = 2 \times 0.0228 = 0.0456.$$

At α = 0.05: since p = 0.0456 < 0.05, we reject H_0. The result is significant — but only just, because the p-value is very close to the significance level. At α = 0.01: since p = 0.0456 > 0.01, we fail to reject H_0. As before, the same data lead to different conclusions at different significance levels. Reporting the actual p-value of 0.046 is far more useful than reporting "significant at 5% but not at 1%", because it gives the reader a precise sense of how strong the evidence really is.

4.12 MEASURING THE POWER OF A HYPOTHESIS TEST

We have so far concentrated almost entirely on the Type I error rate α, the probability of rejecting H_0 when H_0 is true. The other kind of error — failing to reject H_0 when H_0 is false — is equally important, however, and the probability of avoiding it deserves a name of its own. The power of a test is defined as the probability of rejecting H_0 when H_0 is in fact false. In symbols:

$$Power = 1 - \beta = P(reject\ H_0 \mid H_0\ is\ false). \qquad (4.4)$$

A high-power test is one that is very likely to detect a real departure from the null hypothesis when such a departure exists. A low-power test is one that, even if there is a real effect, is unlikely to detect it. The power of a test depends on three things: the level of significance α (a larger α gives higher power, because it widens the critical region); the sample size n (a larger n gives higher power, because it shrinks the standard error); and the size of the true departure from H₀ (a larger departure is easier to detect).

4.12.1 Computing the Power

To compute the power of a test we need to specify a particular value of the parameter under the alternative hypothesis — call it μ_a — and then compute the probability that the test statistic falls in the critical region when the true mean is actually μ_a rather than the null value μ_0. We illustrate the computation with a simple example using the Z-test.

Example 4.9 (Power of a Z-Test)

Consider a right-tailed Z-test of H_0: $\mu = 100$ versus H_1: $\mu > 100$, with σ = 10 and $n = 25$, at the 5% level of significance. (a) Find the critical value of x̄ above which H₀ is rejected. (b) Compute the power of the test if the true mean is in fact $\mu_a = 104$.

Solution.

(a) The standard error under H₀ is σ/√n = 10/5 = 2. For a right-tailed test at α = 0.05, the critical value of Z is $z_{0.05} = 1.645$. The corresponding critical value of x̄ is:

$$\bar{x}_{crit} = \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}} = 100 + 1.645 \cdot 2 = 100 + 3.29 = 103.29.$$

We will reject H₀ if $\bar{x} > 103.29$.

(b) Now suppose the true mean is $\mu_a = 104$. Under this alternative, $\bar{x}$ is normal with mean 104 and standard error 2. The power is the probability that x̄ exceeds 103.29:

$$Power = P(\bar{x} > 103.29 \mid \mu = 104) = P\left(Z > \frac{103.29 - 104}{2}\right) = P(Z > -0.355).$$

From the standard normal table, $P(Z > -0.355) \approx 0.6387$. So the power of the test is approximately 0.64. In words: if the true mean is really 104, this test has only about a 64% chance of detecting that fact and rejecting H₀. The corresponding probability of a Type II error is β = 1 − 0.64 = 0.36, which is uncomfortably high. To increase the power we could either relax α (allowing more Type I errors), increase the sample size, or hope that the true departure is even larger than 4 units.

Example 4.9(b) (Power Increases with Sample Size)

Repeat the power calculation of Example 4.9 with the larger sample size n = 100 (keeping $\sigma = 10, \alpha = 0.05, \mu_a = 104$). Compare the resulting power with the value 0.64 obtained for n = 25, and comment on the effect of increasing the sample size on the power of the test.

Solution. With n = 100, the standard error under H_0 is $\sigma/\sqrt{n} = 10/10 = 1$. The critical value of $\bar{x}$ becomes:

$$\bar{x}_{crit} = 100 + 1.645 \cdot 1 = 101.645.$$

Under the alternative $\mu_a = 104$, $\bar{x}$ is normal with mean 104 and standard error 1. The power is:

$$Power = P(\bar{x} > 101.645 \mid \mu = 104) = P\left(Z > \frac{101.645 - 104}{1}\right) = P(Z > -2.355).$$

From the standard normal table, $P(Z > -2.355) \approx 0.9907$. So the power is approximately 0.99 — almost certainty. By quadrupling the sample size from 25 to 100, we have increased the power of the test from about 64% to about 99%. Equivalently, the probability of a Type II error has dropped from 0.36 to 0.01. This dramatic improvement illustrates one of the most important practical lessons of inferential statistics: when researchers want to detect a real effect that they believe is present, the most reliable way to do so is to ensure that the sample is large enough. Underpowered studies — those with sample sizes too small to detect the effect of interest — are one of the commonest sources of disappointing or inconclusive results in social-science research.

4.12.2 Power and Sample Size

In good research planning, the sample size is chosen so as to give the test adequate power for detecting the smallest departure from H_0 that would be of practical importance. A test with power below about 0.50 has less than an even chance of detecting the effect, and is generally regarded as too weak to be worth running. The conventional minimum target for power in research planning is 0.80 — that is, an 80% chance of detecting the effect. Achieving 80% power for a small effect typically requires a very large sample, which is one of the main reasons that good empirical research is expensive. The detailed mathematics of power calculation and sample-size determination for hypothesis testing is beyond the scope of this introductory course; the interested reader should consult Cohen (1988) for a comprehensive treatment.

4.13 COMMON PITFALLS AND THEIR REMEDIES

Hypothesis testing is one of the most widely used techniques in applied statistics, but it is also one of the most widely misused. Several errors in interpretation are so common that the student should be warned about them explicitly. Avoiding these pitfalls is at least as important as knowing how to compute the test statistics.

4.13.1 Statistical Significance Is Not Practical Importance

A result can be statistically significant — that is, the null hypothesis can be rejected at a small level of significance — without being practically important. With a sufficiently large sample, even a tiny departure from the null hypothesis will be detected. Suppose, for example, that the true mean monthly household expenditure in a city is Rs. 50,005, and the null hypothesis claims it is Rs. 50,000. A sample of one million households would almost certainly reject H_0, because the sample mean would have an extremely small standard error. But the difference of Rs. 5 has no practical importance whatsoever. The researcher must always look at the size of the estimated effect, not just at whether the test rejects, before drawing substantive conclusions.

4.13.2 Failing to Reject Is Not the Same as Accepting

When a hypothesis test fails to reject H_0, the correct phrasing of the conclusion is "we fail to reject H_0" or "the data are consistent with H_0", not "we accept H_0" or "H_0 is true". Failing to reject simply means that the sample evidence is not strong enough to overturn the presumption in favour of H_0; it does not prove that H_0 is correct. The same data may be consistent with many other values of the parameter as well. This is exactly analogous to a "not guilty" verdict in a criminal trial, which means only that the prosecution has not proved guilt — not that the accused is innocent.

4.13.3 Choosing α After Looking at the Data

The level of significance α must be chosen before the data are examined. Choosing α after looking at the data — for example, deciding to "use $\alpha = 0.10$ because that gives a significant result" — completely invalidates the long-run error-control property of the test. The procedure is no longer guaranteed to have a Type I error rate of α; in fact, it has no well-defined Type I error rate at all. Researchers should commit to a value of α (or, better, simply report the p-value) at the planning stage of the study and stick to it.

4.13.4 Multiple Testing

When many hypothesis tests are performed on the same data set, the probability that at least one of them produces a false alarm grows quickly. If 20 independent tests are performed at $\alpha = 0.05$, the probability that at least one of them rejects H_0 purely by chance — even if all 20 null hypotheses are true — is approximately $1 - (0.95)^{20} \approx 0.64$. In other words, in any "fishing expedition" through a large data set, "significant" results are almost guaranteed to appear by chance alone. Various corrections (the Bonferroni correction, the false-discovery-rate procedure, and others) are available to deal with this problem; the simplest is the Bonferroni rule, which divides α by the number of tests being performed. The interested reader is referred to more advanced textbooks for details.

4.13.5 Confusing the p-Value with the Probability that H_0 Is True

The p-value is the probability of obtaining a test statistic at least as extreme as the one observed, computed under the assumption that H_0 is true. It is not the probability that H_0 is true given the data. These two probabilities are mathematically very different (the latter, in fact, requires Bayesian techniques to compute). A p-value of 0.04 does not mean that there is a 4% chance that H_0 is true — it means that, if H_0 were true, there would be only a 4% chance of obtaining the data we actually observed. This subtle but crucial distinction is one of the most commonly misunderstood points in introductory statistics.

4.14 SUMMARY

This unit has developed the theory and practice of one-sample hypothesis testing. The main points may be summarized as follows.

- A statistical hypothesis is a claim about the value of one or more population parameters. Every hypothesis test involves two competing statements: the null hypothesis H_0, which is presumed true unless the evidence contradicts it, and the alternative hypothesis H_1, which is accepted if H_0 is rejected. The two hypotheses must be mutually exclusive and exhaustive.
- A hypothesis test can lead to two kinds of error. A Type I error is rejecting H_0 when H_0 is true; its probability is denoted by α and is also called the level of significance. A Type II error is failing to reject H_0 when H_0 is false; its probability is denoted by β. The power of the test is $1 - \beta$, the probability of correctly rejecting a false H_0.

- A one-tailed test is appropriate when only one direction of departure from H_0 is of interest. A two-tailed test is used when departures in either direction matter. The form of the alternative dictates the form of the critical region.
- The standard six-step procedure is: (1) state the hypotheses, (2) choose α, (3) identify the test statistic and its null distribution, (4) determine the critical region, (5) compute the test statistic from the sample, and (6) decide whether to reject H_0 and state the conclusion.
- For testing a single population mean: when σ is known, the Z-test is used with test statistic $Z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}) \sim N(0,1)$. When σ is unknown, the t-test is used with $t = (\bar{x} - \mu_0)/(s/\sqrt{n}) \sim t(n-1)$. The t-distribution allows for the additional uncertainty caused by estimating σ.
- For testing a single population proportion in a large sample, the test statistic is $Z = (\hat{p} - \pi_0)/\sqrt{[\pi_0(1-\pi_0)/n]} \sim N(0,1)$. The normal approximation requires $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$.
- The p-value of a test is the probability of obtaining a test statistic at least as extreme as the one observed, computed under H_0. The decision rule is "reject H_0 if $p \leq \alpha$". Reporting the p-value is more informative than reporting the bare decision because it lets the reader apply his or her own preferred level of significance.
- The power of a test, $1 - \beta$, depends on α, on n, and on the size of the true departure from H_0. A larger sample, a larger α, or a larger true effect all increase the power. Adequate power (typically at least 0.80) should be ensured by choosing the sample size appropriately at the planning stage of any research study.

In Unit 5 we shall extend exactly the same machinery to two-sample tests, in which two samples are compared with one another in order to draw inferences about the difference between two population means or two population proportions. The student who has thoroughly mastered the one-sample tests of the present unit will find the transition to two-sample tests very natural; the same six-step procedure applies, with only the test statistic and the standard error formula changing.

4.15 SELF-ASSESSMENT EXERCISE

1. Define the null hypothesis and the alternative hypothesis. Why does the null hypothesis always contain an equality sign? What is the role of the null hypothesis in statistical research?
2. Distinguish clearly between a Type I error and a Type II error. Give an example of each in the context of testing a new drug.
3. What is meant by the level of significance α? Why is it usually chosen before the data are collected, rather than after?
4. Distinguish between a one-tailed and a two-tailed test. Give one example of a research question for which each kind of test would be appropriate.
5. Define the p-value of a test. State the decision rule that uses the p-value, and explain why reporting the p-value is more informative than reporting only the decision "reject" or "fail to reject".
6. Explain in detail the standard six-step procedure for carrying out a one-sample hypothesis test. For each step, describe what the researcher does, what concepts are involved, and what could go wrong if the step were skipped or done incorrectly. Illustrate the procedure with a complete numerical example of your own choosing, involving a Z-test of a population mean with σ known.

7. Compare and contrast the Z-test and the t-test for a single population mean. State the assumptions of each, write down the test statistic and its sampling distribution under H_0, and explain when each test should be used. Illustrate the t-test with a complete numerical example.

8. A factory in Sialkot produces footballs whose mean weight is supposed to be 450 grams. The population standard deviation is known to be $\sigma = 12$ grams. A random sample of $n = 36$ footballs gives a mean weight of $\bar{x} = 446$ grams. Test, at the 5% level, whether the mean weight differs from 450 grams. [Answer: $Z = -2.0$; $|Z| > 1.96$; reject H_0]

9. A textile manufacturer claims that the mean breaking strength of his cotton thread is at least 5 kg. The population standard deviation is $\sigma = 0.6$ kg. A random sample of $n = 25$ threads gives a mean of $\bar{x} = 4.78$ kg. Test the manufacturer's claim at $\alpha = 0.05$. [Answer: $Z = -1.833$; left-tailed critical value -1.645; reject H_0]

10. A random sample of $n = 16$ final-year medical students at a university in Lahore gave a mean score of $\bar{x} = 78.5$ marks on a standardized exam, with a sample standard deviation of $s = 6.4$. Test, at $\alpha = 0.05$, whether the population mean differs from the national average of 75 marks. [Answer: $t = 2.188$; $df = 15$; $|t| > 2.131$; reject H_0]

11. A consumer-protection agency draws a random sample of $n = 20$ chocolate bars from a particular brand and finds the mean net weight to be $\bar{x} = 49.2$ grams with a sample standard deviation of $s = 1.8$ grams. The label says the net weight is 50 grams. Test, at $\alpha = 0.01$, whether the mean net weight is less than 50 grams. [Answer: $t = -1.987$; $df = 19$; left-tailed critical value -2.539; fail to reject H_0]

12. A pollster claims that 45% of voters in a constituency intend to vote for a particular party. A random sample of $n = 500$ voters finds 240 in support. Test, at $\alpha = 0.05$, whether the true proportion differs from 0.45. [Answer: $\hat{p} = 0.48$; $Z = 1.349$; $|Z| < 1.96$; fail to reject H_0]

13. A medical researcher claims that fewer than 25% of adults in a certain region have a particular vitamin deficiency. A random sample of $n = 400$ adults finds 88 with the deficiency. Test the claim at $\alpha = 0.05$. [Answer: $\hat{p} = 0.22$; $Z = -1.386$; left-tailed critical value -1.645; fail to reject H_0]

14. A school principal in Peshawar claims that the average IQ of his students exceeds 105. A random sample of $n = 49$ students gives a mean IQ of $\bar{x} = 107.5$ with a sample standard deviation of $s = 9$. Test the claim at $\alpha = 0.05$. [Answer: $t = 1.944$; $df = 48$; right-tailed critical value ≈ 1.677; reject H_0]

15. A large bank claims that the mean waiting time of customers at its branches is no more than 4 minutes. A random sample of $n = 64$ customers gives a mean waiting time of $\bar{x} = 4.45$ minutes with σ known to be 1.2 minutes. Test the claim at $\alpha = 0.01$. [Answer: $Z = 3.0$; right-tailed critical value 2.33; reject H_0]

16. Consider a right-tailed Z-test of H_0: $\mu = 50$ versus H_1: $\mu > 50$ at $\alpha = 0.05$, with $\sigma = 8$ and $n = 64$. (a) Find the critical value of $\bar{x}$. (b) Compute the power of the test if the true mean is in fact $\mu_a = 52$. [Answers: (a) $\bar{x}_{crit} \approx 51.645$; (b) Power ≈ 0.6368]

17. In a hypothesis test, the test statistic Z came out to be 2.10. (a) Compute the p-value if the test is two-tailed. (b) Compute the p-value if the test is right-tailed. (c) Would you reject H_0 at $\alpha = 0.05$ in each case? [Answers: (a) $p \approx 0.0358$, reject; (b) $p \approx 0.0179$, reject; both reject at $\alpha = 0.05$]

UNIT 5: TESTING HYPOTHESES: TWO SAMPLE TESTS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Distinguish clearly between independent and dependent (paired) samples, and select the appropriate test for each kind of data.
2. Carry out a two-sample test for the difference between two population means, using either the large-sample Z-test, the pooled-variance t-test, the Welch unpooled t-test, or the paired-difference t-test, and explain the assumptions of each.
3. Carry out a two-sample test for the difference between two population proportions in large samples, using both the unpooled and the pooled standard-error formulas, and recognize when each is appropriate.
4. Compute and interpret a p-value for a two-sample test, and use it to make decisions at any chosen level of significance.
5. Identify the most common pitfalls in two-sample inference — confusing independence with paired samples, ignoring unequal variances, comparing samples that come from different populations — and avoid them in their own research.

5.1 INTRODUCTION

In Unit 4 we developed the formal machinery of one-sample hypothesis testing. We learned how to formulate the null and alternative hypotheses, how to choose a level of significance, how to construct critical regions and to compute test statistics, and how to interpret the resulting decisions in the context of a research problem. Throughout that unit, however, we considered only the simplest situation: a single sample drawn from a single population, with the goal of testing a claim about a single population parameter. In real research, this situation — though not uncommon — is the exception rather than the rule. Most empirical investigations in the social sciences are comparative in nature: they contrast two groups, two treatments, two cities, two periods of time, or two methods, and the researcher's central question is whether the two groups differ from one another in some specified way.

The present unit develops the inferential techniques needed to answer such comparative questions. We shall see that, with only minor modifications, the same six-step procedure that we used for one-sample tests in Unit 4 can be applied directly to two-sample tests. The hypotheses now refer to the difference between two population parameters; the test statistic is constructed from the difference between two sample statistics; and the standard error formula is correspondingly more elaborate, because it has to combine the variability of two samples into a single number. Apart from these changes, the underlying logic is exactly the same as before, and the student who has thoroughly absorbed Unit 4 will find this unit a natural continuation rather than a fresh start.

There is one important new idea in this unit that has no analogue in the one-sample case, and that is the distinction between independent samples and dependent (or paired) samples. When the two samples consist of completely separate units — for example, fifty male students and fifty female students drawn at random from a university — the samples are said to be independent, and the standard two-sample tests apply. When the two samples consist of matched pairs of units — for example, the blood pressure of fifty patients before and after taking a new medicine — the samples are said to be dependent, and a quite different procedure (the paired-difference t-test) must be used. Failure to recognize the difference between these two situations is one of the commonest errors in applied statistical work, and we shall therefore devote considerable attention to it.

The unit is organized as follows. Section 5.2 develops the distinction between independent and paired samples in detail. Section 5.3 derives the sampling distribution of the difference between two sample means, which is the theoretical foundation of all the two-sample tests for means. Sections 5.4 to 5.7 then take up the four main two-sample tests for means, in order of increasing subtlety: the large-sample Z-test, the pooled-variance t-test, the Welch unpooled t-test, and the paired-difference t-test. Section 5.8 develops the corresponding tests for two population proportions. Sections 5.9 and 5.10 give the p-value version of the procedure and a decision map for choosing the right test. Section 5.11 discusses the most common pitfalls. The unit closes with a summary, a generously extended self-assessment exercise containing both routine and conceptual numerical problems, and the usual list of references.

The level of difficulty in this unit is somewhat higher than in the previous ones. The student is now expected not only to apply the formulas mechanically, but also to make conscious choices about which formula is appropriate in any given situation, and to defend those choices in writing. The exercises at the end of the unit reflect this raised expectation: many of them require the student first to identify the correct test before performing the computation.

5.2 INDEPENDENT VS. DEPENDENT (PAIRED) SAMPLES

The first decision the researcher must make in any two-sample problem is whether the two samples are independent or dependent. This decision is not a matter of mathematical convenience; it is a matter of how the data were actually collected, and an incorrect classification will lead to the wrong test being applied — sometimes with seriously misleading consequences.

5.2.1 Independent Samples

Two samples are said to be independent (sometimes called unrelated) if the units in one sample have nothing to do with the units in the other sample. Knowing the value of any observation in the first sample tells us nothing about the value of any observation in the second sample. The two samples are typically drawn separately from two different populations, or from two different subgroups of the same population, and the units in the two samples are physically distinct individuals or objects.

Examples of independent samples in social-science research include the following:

1. Sample 1: 40 male students randomly selected from BS Economics. Sample 2: 35 female students randomly selected from BS Economics. The 40 men and the 35 women are different people.
2. Sample 1: 60 households from urban Lahore. Sample 2: 60 households from rural Punjab. The two samples come from two different sub-populations.
3. Sample 1: 25 patients given a new medicine. Sample 2: 25 different patients given a placebo. The two groups consist of separate individuals.
4. Sample 1: A random sample of monthly sales in 2024. Sample 2: A random sample of monthly sales in 2025. (Provided the months chosen in the two years are not the same calendar months — otherwise the samples would be paired.)

Independent samples need not be of the same size. The two sample sizes are usually denoted n_1 and n_2, and they may be equal or unequal. The two samples are typically analyzed by computing two separate sample means ($\bar{x}_1$ and $\bar{x}_2$) and two separate sample standard deviations (s_1 and s_2), and then combining them in the appropriate two-sample test statistic.

5.2.2 Dependent (Paired) Samples

Two samples are said to be dependent or paired (sometimes called matched, related, or repeated-measures samples) when each observation in one sample is naturally linked to a specific observation in the other sample. The pairing may arise in any of several different ways, but the essential feature is that the two samples are the same size and that the i-th observation of the first sample "belongs together" with the i-th observation of the second sample. Some common situations in which paired samples arise are:

1. Before-and-after measurements on the same units. For example, the blood pressure of 30 patients measured before and again after taking a new medicine for two weeks. Each patient produces two measurements, one in the "before" sample and one in the "after" sample.
2. Matched pairs by design. For example, in a study comparing two teaching methods, each student in the first group is matched with a student in the second group of the same age, sex, IQ and family background, so that any differences observed between the two groups can be attributed to the teaching method rather than to the personal characteristics of the students.
3. Repeated measurements on the same units under two conditions. For example, the reaction times of 20 drivers measured first in normal lighting and then again in dim lighting. Each driver contributes two reaction times, one under each condition.

In all of these situations, the two samples are linked by a one-to-one correspondence: it makes sense to say "this observation in sample 1 corresponds to that observation in sample 2". The two samples must therefore be of the same size ($n_1 = n_2 = n$), and the analysis must take the pairing into account.

5.2.3 Why the Distinction Matters

A natural question at this point is: why does it matter whether the samples are independent or paired? The answer lies in the variability of the test statistic. When samples are paired, the differences within pairs typically have much smaller variability than the individual measurements themselves, because much of the variability between individuals (their age, sex, baseline level, family background, and so on) cancels out when we look at the within-pair difference. The standard error of the estimated treatment effect is therefore smaller in a paired design than in an independent-samples design of the same total size. In other words, paired designs are typically more powerful than independent-samples designs at detecting a real effect.

Treating paired data as if they were independent throws away this advantage and gives an unnecessarily wide standard error, leading to lower power and possibly to a failure to detect a real effect. Conversely, treating independent samples as if they were paired is technically incorrect (because there is no genuine pairing) and leads to invalid inferences. In short: the choice of test must reflect the actual structure of the data, and the researcher must make this choice consciously and correctly. We shall return to this point in Section 5.7 with a worked example showing exactly how much can be gained by recognizing a paired structure correctly.

5.3 THE SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO MEANS

Before we can construct a hypothesis test about the difference between two population means, we need to know the sampling distribution of the corresponding sample statistic, namely $(\bar{x}_1 - \bar{x}_2)$. The derivation of this sampling distribution follows directly from the theory developed in Unit 2, and we present it in some detail because it is the foundation of every two-sample test for means in this unit.

Suppose we have two independent random samples of sizes n_1 and n_2 drawn from two populations whose means are μ_1 and μ_2 and whose standard deviations are σ_1 and σ_2 respectively. Let $\bar{x}_1$ and $\bar{x}_2$ denote the two sample means. Two facts from Unit 2 will be needed:

1. The expected value of the difference of two independent random variables is the difference of their expected values: $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$.
2. The variance of the difference of two independent random variables is the sum (not the difference) of their variances: $\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \mathrm{Var}(\bar{x}_1) + \mathrm{Var}(\bar{x}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$.

The plus sign in the second formula is sometimes a stumbling block for beginning students. The variances add even though the means subtract because variances measure spread around the mean, and spread is positive whether we are adding or subtracting. The standard error of the difference is the square root of the sum:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5.1)$$

Notice that this is exactly the same formula we met in Section 3.8 in connection with confidence intervals for differences of means; the same standard error appears in both estimation and hypothesis testing. By the Central Limit Theorem, when both n_1 and n_2 are sufficiently large (each at least 30, say), the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is approximately normal:

$$(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \qquad (5.2)$$

and the standardized statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (5.3)$$

follows the standard normal distribution N(0, 1). This is the test statistic for the large-sample Z-test of Section 5.4. When the population variances are unknown and the samples are small, we shall need to estimate σ_1^2 and σ_2^2 from the samples, and the standard normal distribution will be replaced by Student's t-distribution with an appropriate number of degrees of freedom — exactly as in the one-sample case. The details depend on whether or not we are willing to assume that the two population variances are equal, and we take up the two cases separately in Sections 5.5 and 5.6.

5.4 TESTING THE DIFFERENCE BETWEEN TWO MEANS: LARGE SAMPLES (Z-TEST)

We begin with the simplest two-sample test: the test for the difference between two means when both samples are large ($n_1 \geq 30$ and $n_2 \geq 30$). When the samples are large, the Central Limit Theorem guarantees that the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is approximately normal regardless of the shape of the underlying populations, and we can use the standard normal table to determine the critical region. Furthermore, in large samples the sample standard deviations s_1 and s_2 are very close to their respective population values σ_1 and σ_2, so we may safely substitute the sample values for the unknown population values in the standard error formula.

5.4.1 The Hypotheses

We are typically interested in testing one of the following three pairs of hypotheses:

1. Two-tailed: $H_0: \mu_1 - \mu_2 = \delta_0$ versus $H_1: \mu_1 - \mu_2 \neq \delta_0$.
2. Right-tailed: $H_0: \mu_1 - \mu_2 = \delta_0$ versus $H_1: \mu_1 - \mu_2 > \delta_0$.
3. Left-tailed: $H_0: \mu_1 - \mu_2 = \delta_0$ versus $H_1: \mu_1 - \mu_2 < \delta_0$.

Here δ_0 is some specified hypothesized value of the difference between the two means. In the most common case the hypothesized difference is zero — that is, the null hypothesis says that the two population means are equal — and the formulas simplify accordingly. We shall therefore present the formulas with $\delta_0 = 0$ from now on; the extension to non-zero δ_0 is straightforward and is illustrated in one of the exercises at the end of the unit.

5.4.2 The Test Statistic

Under $H_0: \mu_1 = \mu_2$, equation (5.3) reduces to:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (5.4)$$

which follows the standard normal distribution N(0, 1) under H_0. When σ_1 and σ_2 are unknown but n_1 and n_2 are both large (at least 30), we replace σ_1 by s_1 and σ_2 by s_2, and the resulting statistic still follows N(0, 1)

approximately. The form of the critical region depends on the alternative hypothesis exactly as in Section 4.8: for a two-tailed test we reject H_0 if $|Z| > z_{\alpha/2}$, and so on for the one-tailed cases.

Example 5.1 (Z-Test for Two Means, Two-tailed)

A consumer-protection agency wants to compare the mean fat content of a particular brand of biscuits manufactured at two different factories — one in Lahore and one in Karachi. A random sample of $n_1 = 50$ biscuits from the Lahore factory gives a mean fat content of $\bar{x}_1 = 5.4$ grams with a sample standard deviation of $s_1 = 0.6$ grams, while an independent random sample of $n_2 = 60$ biscuits from the Karachi factory gives $\bar{x}_2 = 5.1$ grams with $s_2 = 0.7$ grams. Test, at the 5% level of significance, whether the mean fat content differs between the two factories.

Solution.

Step 1. Hypotheses. We are interested in any difference in either direction:

$$H_0: \mu_1 - \mu_2 = 0 \quad versus \quad H_1: \mu_1 - \mu_2 \neq 0 \quad (two-tailed\ test).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic. Both samples are large ($n_1 = 50$, $n_2 = 60$), so we use the Z-test (5.4) with sample standard deviations replacing the unknown population standard deviations.

Step 4. Critical region. For a two-tailed test at $\alpha = 0.05$, the critical values are ± 1.96. We reject H_0 if $|Z| > 1.96$.

Step 5. Computation. The standard error is:

$$SE = \sqrt{\frac{(0.6)^2}{50} + \frac{(0.7)^2}{60}} = \sqrt{0.0072 + 0.008167} = \sqrt{0.015367} = 0.1240.$$

$$Z = \frac{5.4 - 5.1}{0.1240} = \frac{0.30}{0.1240} = 2.419.$$

Step 6. Decision and conclusion. Since $|Z| = 2.419 > 1.96$, we reject H_0 at the 5% level of significance. There is sufficient evidence to conclude that the mean fat content of biscuits differs between the Lahore and Karachi factories. The Lahore biscuits appear to have a slightly higher mean fat content (by about 0.3 grams per biscuit). Notice also that the test statistic exceeds 1.96 but is well below 2.576 (the 1% two-tailed critical value), so the result is significant at the 5% level but not at the 1% level — a useful illustration of how the choice of α affects the conclusion.

5.5 TESTING THE DIFFERENCE BETWEEN TWO MEANS: SMALL SAMPLES, EQUAL VARIANCES (POOLED t-TEST)

When the samples are small (one or both sample sizes below 30) and the population standard deviations are unknown, we cannot rely on the Central Limit Theorem to justify the use of the standard normal distribution. We must therefore make additional assumptions about the populations from which the samples are drawn, and we must use the Student's t-distribution rather than the standard normal in setting the critical values. The pooled-variance t-test, which we develop in this section, makes the additional assumption that the two

populations have the same (unknown) standard deviation; the Welch t-test of the next section relaxes this assumption.

5.5.1 The Assumptions

The pooled-variance t-test of H_0: $\mu_1 = \mu_2$ rests on three assumptions, all of which must be reasonably satisfied for the procedure to be valid:

1. The two samples are independent random samples drawn from their respective populations.
2. Both populations are normal (or at least approximately so). For modest departures from normality, the test is fairly robust, especially when the two sample sizes are similar; for very skewed populations, however, the test can give misleading results when the samples are small.
3. The two populations have the same standard deviation: $\sigma_1 = \sigma_2 = \sigma$. This is the so-called assumption of equal variances or homoscedasticity. It is the most restrictive of the three assumptions, and the validity of the test depends critically on it. We discuss how to check this assumption — and what to do if it fails — in Section 5.6.

When all three assumptions are satisfied, the pooled-variance t-test is the most powerful procedure available for comparing two population means in small samples. When the equal-variances assumption is doubtful, the Welch t-test of Section 5.6 is the safer choice.

5.5.2 The Pooled Variance and the Test Statistic

Under the assumption $\sigma_1 = \sigma_2 = \sigma$, the obvious thing to do is to combine the information in both samples to estimate the single common variance σ^2. The combined estimator, called the pooled sample variance, is the weighted average of the two sample variances, with weights given by the corresponding degrees of freedom $(n_1 - 1)$ and $(n_2 - 1)$:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (5.5)$$

The pooled sample variance is itself an unbiased estimator of the common population variance σ^2 (provided the assumption $\sigma_1 = \sigma_2$ is correct). It always falls between the two individual sample variances s_1^2 and s_2^2, and its weighting reflects the fact that the larger sample contributes more information to the estimate. The square root s_p is the corresponding pooled sample standard deviation.

The test statistic for testing H_0: $\mu_1 = \mu_2$ is then:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (5.6)$$

which follows Student's t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom under H_0. The form of the critical region depends on the alternative hypothesis in the usual way: $|t| > t_{\alpha/2,\, df}$ for a two-tailed test, $t > t_{\alpha,\, df}$ for a right-tailed test, and $t < -t_{\alpha,\, df}$ for a left-tailed test.

Example 5.2 (Pooled t-Test, Two-tailed)

A study was conducted in Punjab to compare the mean monthly food expenditure of urban and rural households. A random sample of $n_1 = 12$ urban households gave a mean of $\bar{x}_1 = 22{,}400$ rupees with a sample standard deviation of $s_1 = 3{,}200$ rupees, and an independent random sample of $n_2 = 14$

rural households gave a mean of $\bar{x}_2 = 17{,}800$ rupees with a sample standard deviation of $s_2 = 2{,}900$ rupees. Assume that the two populations are approximately normal with equal variances. Test, at the 5% level of significance, whether the mean monthly food expenditure differs between urban and rural households.

Solution.

Step 1. Hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \quad versus \quad H_1: \mu_1 - \mu_2 \neq 0 \quad (two-tailed).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: pooled-variance t-test (5.6), with $df = n_1 + n_2 - 2 = 12 + 14 - 2 = 24$.

Step 4. Critical region. For $df = 24$ at $\alpha = 0.05$, two-tailed, $t_{0.025,\,24} = 2.064$. We reject H_0 if $|t| > 2.064$.

Step 5. Computation of the pooled variance:

$$s_p^2 = \frac{(11)(3200)^2 + (13)(2900)^2}{24}$$

$$= \frac{11(10{,}240{,}000) + 13(8{,}410{,}000)}{24} = \frac{112{,}640{,}000 + 109{,}330{,}000}{24} = \frac{221{,}970{,}000}{24}$$
$$= 9{,}248{,}750.$$

$$s_p = \sqrt{9{,}248{,}750} \approx 3041.18 \, rupees.$$

Now compute the test statistic:

$$t = \frac{22400 - 17800}{3041.18 \cdot \sqrt{\frac{1}{12} + \frac{1}{14}}}$$

$$= \frac{4600}{3041.18 \cdot \sqrt{0.0833 + 0.0714}} = \frac{4600}{3041.18 \cdot 0.3934} = \frac{4600}{1196.40} = 3.846.$$

Step 6. Decision. Since $|t| = 3.846 > 2.064$, we reject H_0 at the 5% level. In fact the test statistic also far exceeds the 1% critical value $t_{0.005,\,24} = 2.797$ and the 0.1% critical value $t_{0.005,\,24} \approx 3.745$, so the result is highly significant. We conclude that there is very strong evidence that the mean monthly food expenditure of urban households in Punjab differs from that of rural households. The point estimate of the difference is 22,400 − 17,800 = Rs. 4,600 in favour of the urban households, which is large enough to be of practical importance and not just statistical significance.

5.6 TESTING THE DIFFERENCE BETWEEN TWO MEANS: SMALL SAMPLES, UNEQUAL VARIANCES (WELCH'S t-TEST)

The pooled-variance t-test of Section 5.5 makes the strong assumption that the two populations have the same standard deviation. In many practical situations this assumption is doubtful — for example, when the two populations are known on substantive grounds to differ in variability, or when the two sample standard deviations are markedly different from one another. A useful rule of thumb is that the equal-variances assumption is acceptable provided the larger sample standard deviation is not more than about twice the

smaller one; if the ratio s_1/s_2 (or its reciprocal) exceeds 2, the assumption is suspect and the pooled t-test should not be used. A formal F-test for equal variances also exists; we shall meet it in Unit 6.

When the equal-variances assumption fails, the appropriate procedure is the Welch unpooled t-test, named after the British statistician Bernard Lewis Welch who proposed it in 1947. Welch's test does not pool the two sample variances; instead, it uses each sample variance separately in the standard error formula. This makes the procedure valid even when the two population variances differ from one another. The price of this added flexibility is a slightly more complicated formula for the degrees of freedom of the t-distribution, which is no longer simply $n_1 + n_2 - 2$.

5.6.1 The Test Statistic

The Welch test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (5.7)$$

and follows Student's t-distribution approximately with degrees of freedom given by the so-called Welch–Satterthwaite formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad (5.8)$$

In practice this formula gives a fractional number of degrees of freedom, which is then rounded down to the nearest whole number for the purpose of looking up critical values in the t-table. (Some texts and software round to the nearest integer instead; the difference is rarely large.) For all but the smallest samples, the Welch degrees of freedom turn out to be only slightly smaller than the corresponding pooled degrees of freedom ($n_1 + n_2 - 2$), so the Welch test is only slightly less powerful than the pooled test when the equal-variances assumption is correct, but it is much safer when that assumption fails.

Example 5.3 (Welch's t-Test)

A pharmacology researcher in Islamabad wants to compare the mean reaction time (in milliseconds) of patients treated with two different anti-anxiety medicines. A random sample of $n_1 = 10$ patients given Medicine A gave a mean reaction time of $\bar{x}_1 = 320$ ms with a sample standard deviation of $s_1 = 25$ ms. An independent random sample of $n_2 = 12$ patients given Medicine B gave a mean reaction time of $\bar{x}_2 = 295$ ms with a sample standard deviation of $s_2 = 60$ ms. Test, at the 5% level of significance, whether the two medicines produce different mean reaction times.

Solution. Notice that the two sample standard deviations are very different from one another (the ratio is 60/25 = 2.4, which exceeds the rule of thumb of 2). The equal-variances assumption is therefore suspect, and we should use the Welch unpooled t-test rather than the pooled-variance t-test.

Step 1. Hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \quad versus \quad H_1: \mu_1 - \mu_2 \neq 0 \quad (two-tailed).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: Welch unpooled t-test (5.7).

Step 4. Computing the Welch degrees of freedom. Let $A = s_1^2/n_1 = 625/10 = 62.5$ and $B = s_2^2/n_2 = 3600/12 = 300$. Then:

$$df = \frac{(A + B)^2}{\frac{(A)^2}{n_1 - 1} + \frac{(B)^2}{n_2 - 1}} = \frac{(362.5)^2}{\frac{3906.25}{9} + \frac{90000}{11}}$$

$$= \frac{131{,}406.25}{434.03 + 8181.82} = \frac{131{,}406.25}{8615.85} \approx 15.25.$$

Rounding down, we use df = 15. From the t-table, $t_{0.025,\,15} = 2.131$. We reject H_0 if $|t| > 2.131$.

Step 5. Computation of the test statistic:

$$t = \frac{320 - 295}{\sqrt{62.5 + 300}} = \frac{25}{\sqrt{362.5}} = \frac{25}{19.04} = 1.313.$$

Step 6. Decision. Since $|t| = 1.313 < 2.131$, we fail to reject H_0 at the 5% level. There is not sufficient evidence to conclude that the two medicines produce different mean reaction times. The observed difference of 25 ms is well within what could be expected by chance alone, given the variability in the data. A larger study would be needed to determine whether a real difference exists. Notice that, had we (incorrectly) used the pooled t-test for this problem, the test statistic and the conclusion would have been somewhat different. The student is invited to repeat the calculation under the pooled assumption as an exercise, and to note how the choice of test affects the result.

5.7 TESTING THE DIFFERENCE BETWEEN TWO MEANS WITH DEPENDENT (PAIRED) SAMPLES

We now turn to the case of paired samples. As we explained in Section 5.2, the two samples here are linked by a one-to-one correspondence: the i-th observation in the first sample "belongs together" with the i-th observation in the second sample, for i = 1, 2, …, n. Examples include before-and-after measurements on the same individuals, matched pairs of similar individuals assigned to two treatments, and repeated measurements on the same units under two different conditions.

The key idea behind the paired-difference t-test is very simple: instead of analyzing the two samples separately, we compute the within-pair differences $d_i = x_{1i} - x_{2i}$ for i = 1, 2, …, n, and then we analyze the resulting single sample of n differences using a one-sample t-test. The hypothesis "the two population means are equal" becomes "the mean of the differences in the population is zero", and the procedure of Section 4.9 applies directly. This conversion of a two-sample problem into a one-sample problem is one of the most elegant tricks in inferential statistics.

5.7.1 The Test Statistic

Let $d_i = x_{1i} - x_{2i}$ (i = 1, 2, …, n) denote the n within-pair differences. Compute the sample mean and the sample standard deviation of these differences:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}, \quad s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n - 1}} \qquad (5.9)$$

Then, to test H_0: $\mu_d = 0$ against an appropriate alternative, the test statistic is:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad (5.10)$$

which follows Student's t-distribution with df = n − 1 degrees of freedom under H_0. Notice carefully: the degrees of freedom are n − 1, not 2n − 2 as in the pooled two-sample t-test, because we have n pairs and therefore effectively n observations on the difference variable. This is an important practical advantage of paired designs: they "spend" only n − 1 degrees of freedom for n pairs of observations, rather than 2n − 2.

Example 5.4 (Paired t-Test — Before and After)

A pharmacologist at a hospital in Karachi wants to test whether a new diet program reduces the systolic blood pressure of overweight patients. He randomly selects 10 patients, measures their systolic blood pressure before the program begins, and again after the program has been running for 12 weeks. The data (in mmHg) are as follows:

Patient	1	2	3	4	5	6	7	8	9	10
Before (x_1)	155	142	160	148	170	152	165	158	146	163
After (x_2)	148	140	152	145	162	148	160	152	142	156

Test, at the 5% level of significance, whether the diet program reduces the mean systolic blood pressure.

Solution. The two samples are clearly paired (each patient gives a "before" and an "after" reading), so we use the paired-difference t-test.

Step 1. Let $d = x_1 - x_2$ be the within-patient reduction. The hypotheses are:

$$H_0: \mu_d = 0 \quad versus \quad H_1: \mu_d > 0 \quad (right - tailed\ test).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: paired-difference t-test (5.10), with df = n − 1 = 9.

Step 4. Critical region. For a right-tailed test at $\alpha = 0.05$ with df = 9, $t_{0.05,\,9} = 1.833$. We reject H_0 if $t > 1.833$.

Step 5. Computation of the differences and their sample mean and standard deviation:

Patient	1	2	3	4	5	6	7	8	9	10
$d = x_1 - x_2$	7	2	8	3	8	4	5	6	4	7

The sum of the differences is 7+2+8+3+8+4+5+6+4+7 = 54, so:

$$\bar{d} = \frac{54}{10} = 5.4\ mmHg.$$

The squared deviations from $\bar{d} = 5.4$ are:

$(7-5.4)^2 = 2.56$; $(2-5.4)^2 = 11.56$; $(8-5.4)^2 = 6.76$; $(3-5.4)^2 = 5.76$; $(8-5.4)^2 = 6.76$; $(4-5.4)^2 = 1.96$; $(5-5.4)^2 = 0.16$; $(6-5.4)^2 = 0.36$; $(4-5.4)^2 = 1.96$; $(7-5.4)^2 = 2.56$. Sum = 40.40.

$$s_d^2 = \frac{40.40}{9} = 4.489 \quad \Rightarrow \quad s_d = \sqrt{4.489} = 2.119\ mmHg.$$

The test statistic is:

$$t = \frac{5.4}{2.119/\sqrt{10}} = \frac{5.4}{2.119/3.162} = \frac{5.4}{0.6701} = 8.060.$$

Step 6. Decision and conclusion. Since t = 8.060 ≫ 1.833, we reject H_0 at the 5% level — and indeed at any conventional level (the result is significant at α = 0.001 as well, since ($t_{0.001,\,9} \approx 4.297$). There is overwhelming evidence that the diet program reduces the mean systolic blood pressure of overweight patients. The estimated reduction is 5.4 mmHg, with a tightly clustered set of within-patient differences. This is exactly the kind of result that would have been much weaker — possibly even non-significant — if the same data had been (incorrectly) treated as two independent samples, because the much larger between-patient variability would have inflated the standard error. The student is invited to verify this by re-analyzing the same data as if the "before" and "after" measurements came from two independent samples.

5.8 TESTING THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS

Just as we may wish to compare two population means, we may wish to compare two population proportions. Has the proportion of literate adults in a district increased between two surveys? Does a new advertising campaign produce a higher recognition rate than the old one? Is the proportion of women in the workforce higher in urban Punjab than in rural Sindh? In each case the parameter of interest is the difference ($\pi_1 - \pi_2$), and the data consist of two sample proportions $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$ computed from two independent random samples.

5.8.1 The Sampling Distribution

From Section 2.15 we know that, for sufficiently large samples, $\hat{p}_1$ is approximately normal with mean π_1 and variance $\pi_1(1-\pi_1)/n_1$, and similarly $\hat{p}_2$ is approximately normal with mean π_2 and variance $\pi_2(1-\pi_2)/n_2$. Because the two samples are independent, the difference $(\hat{p}_1 - \hat{p}_2)$ is also approximately normal, with mean $(\pi_1 - \pi_2)$ and variance equal to the sum of the two individual variances:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \qquad (5.11)$$

The conventional rule of thumb for the normal approximation to apply is that all four products $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ should be at least 5 (and preferably at least 10).

5.8.2 The Pooled and Unpooled Standard Errors

There is a subtlety here that does not arise in the corresponding case for two means, and the student should pay close attention to it. When we set up the null hypothesis H_0: $\pi_1 = \pi_2$, we are assuming that the two populations share a single common proportion, call it π. Under H_0, the two unknown proportions in the variance formula (5.11) are equal to this common π, and the variance becomes:

$$Var(\hat{p}_1 - \hat{p}_2) \mid H_0 = \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad (5.12)$$

The unknown common proportion π is then estimated from the combined sample by the pooled sample proportion:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \qquad (5.13)$$

which is simply the total number of "successes" in the two samples combined, divided by the total sample size. Substituting $\hat{p}$ for π in the variance formula gives the pooled standard error:

$$SE_{pooled} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad (5.14)$$

and the test statistic for testing H_0: $\pi_1 = \pi_2$ is:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{pooled}} \qquad (5.15)$$

which follows the standard normal distribution N(0, 1) under H_0. The pooled formula is the correct one to use when testing H_0: $\pi_1 = \pi_2$, because under the null hypothesis the two populations really do share a common proportion, and pooling gives a more precise estimate of that common value.

A subtle but important contrast. When constructing a confidence interval for $(\pi_1 - \pi_2)$, as we did in Section 3.10, we do NOT pool, because there is no null hypothesis to justify treating the two proportions as equal. The unpooled standard error of equation (5.11), with $\hat{p}_1$ and $\hat{p}_2$ substituted for the unknown population values, is then the correct one. So:

1. For testing H_0: $\pi_1 = \pi_2$, use the pooled standard error.
2. For constructing a confidence interval for $(\pi_1 - \pi_2)$, use the unpooled standard error.

This is one of the rare situations in introductory statistics where the same parameter is estimated by two different formulas in two different contexts, and the student should make a careful note of it.

Example 5.5 (Test for Two Proportions, Pooled)

A market researcher wishes to compare the proportion of households that own a microwave oven in two cities of Pakistan. A random sample of $n_1 = 400$ households in Lahore finds 152 own a microwave, while an independent random sample of $n_2 = 350$ households in Faisalabad finds 105 own a microwave. Test, at the 5% level of significance, whether the proportion of microwave-owning households differs between the two cities.

Solution.

Step 1. Hypotheses:

$$H_0: \pi_1 - \pi_2 = 0 \quad versus \quad H_1: \pi_1 - \pi_2 \neq 0 \quad (two-tailed).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: pooled-proportion Z-test (5.15).

Step 4. Critical region: $|Z| > 1.96$.

Step 5. Computation. The two sample proportions are:

$$\hat{p}_1 = \frac{152}{400} = 0.380 \;,\; \hat{p}_2 = \frac{105}{350} = 0.300.$$

The pooled sample proportion is:

$$\hat{p} = \frac{152 + 105}{400 + 350} = \frac{257}{750} = 0.3427.$$

The pooled standard error is:

$$SE_{pooled} = \sqrt{(0.3427)(0.6573)\left(\frac{1}{400} + \frac{1}{350}\right)}$$

$$= \sqrt{(0.2253)(0.0025 + 0.002857)} = \sqrt{(0.2253)(0.005357)} = \sqrt{0.001207} = 0.03474.$$

$$Z = \frac{0.380 - 0.300}{0.03474} = \frac{0.080}{0.03474} = 2.303.$$

Step 6. Decision and conclusion. Since |Z| = 2.303 > 1.96, we reject H_0 at the 5% level. There is sufficient evidence to conclude that the proportion of households owning a microwave oven differs between Lahore and Faisalabad. The point estimate suggests that microwave ownership is about 8 percentage points higher in Lahore (38%) than in Faisalabad (30%). At the 1% level (critical value 2.576), the result would have been just non-significant — another illustration of how the strength of the conclusion depends on the chosen α.

Example 5.5(b) (Test for Two Proportions, One-tailed)

A government health program is launched in a particular district of Sindh with the aim of increasing the proportion of children fully immunised against polio. A baseline survey one year before the program found that 380 out of $n_1 = 500$ children were fully immunised. A follow-up survey one year after the program found that 462 out of $n_2 = 550$ children were fully immunised. Test, at the 5% level of significance, whether the program has succeeded in raising the immunisation rate. (Note that the two samples are independent because they consist of different children at different times.)

Solution.

Step 1. Let π_1 = baseline immunisation proportion and π_2 = follow-up immunisation proportion. The program was meant to raise the rate, so we wish to detect $\pi_2 > \pi_1$, i.e. $\pi_1 - \pi_2 < 0$:

$$H_0: \pi_1 - \pi_2 = 0 \quad versus \quad H_1: \pi_1 - \pi_2 < 0 \quad (left - tailed).$$

Step 2. α = 0.05.

Step 3. Test statistic: pooled-proportion Z-test.

Step 4. Critical region. For a left-tailed test at α = 0.05, the critical value is $-z_{0.05} = -1.645$. We reject H_0 if $Z < -1.645$.

Step 5. Computation:

$$\hat{p}_1 = \frac{380}{500} = 0.760 \ , \ \hat{p}_2 = \frac{462}{550} = 0.840.$$

$$\hat{p} = \frac{380 + 462}{500 + 550} = \frac{842}{1050} = 0.8019.$$

$$SE_{pooled} = \sqrt{(0.8019)(0.1981)\left(\frac{1}{500} + \frac{1}{550}\right)}$$

$$= \sqrt{(0.1589)(0.002 + 0.001818)} = \sqrt{(0.1589)(0.003818)} = \sqrt{0.000607} = 0.02463.$$

$$Z = \frac{0.760 - 0.840}{0.02463} = \frac{-0.080}{0.02463} = -3.249.$$

Step 6. Decision and conclusion. Since $Z = -3.249 < -1.645$, we reject H_0 at the 5% level. In fact $Z < -2.576 = -z_{0.005}$ as well, so the result is significant even at the 1% level. There is very strong evidence that the immunisation program has succeeded in raising the proportion of fully immunised children. The estimated increase is $0.840 - 0.760 = 0.08$, that is, an 8-percentage-point gain — a result that is both statistically significant and substantively important from a public-health point of view.

5.9 THE p-VALUE REVISITED AND THE TWO-SAMPLE DECISION RULE

The p-value approach that we developed in Section 4.11 applies to two-sample tests in exactly the same way as to one-sample tests. After computing the test statistic (Z, t, or any other), we look up the area in the appropriate tail (or both tails, for a two-tailed test) of the relevant null distribution. The decision rule is unchanged: reject H_0 if $p \le \alpha$, and fail to reject H_0 if $p > \alpha$. The interpretation of the p-value is also unchanged: it is the probability, computed under the assumption that H_0 is true, of obtaining a test statistic at least as extreme as the one observed.

The advantage of reporting the p-value rather than just the binary decision is the same as before: it lets the reader apply his or her own preferred level of significance, and it conveys a continuous measure of the strength of the evidence rather than a coarse yes-or-no answer. In modern empirical research, two-sample test results are almost always reported with their p-values, both in academic journals and in the output of statistical software packages.

Example 5.6 (Computing a p-Value for a Two-Sample Test)

In Example 5.2 (the urban-versus-rural food expenditure study), we computed a pooled-variance t-test statistic of $t = 3.846$ with $df = 24$ in a two-tailed test. (a) Use the t-table to bracket the p-value. (b) State the conclusion at $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.001$.

Solution.

(a) From the t-table with $df = 24$, the relevant critical values are $t_{0.005,\,24} = 2.797$ and $t_{0.0005,\,24} \approx 3.745$. Our observed $t = 3.846$ exceeds even the latter, so the upper-tail probability beyond 3.846 is less than 0.0005, and the two-tailed p-value is less than $2 \times 0.0005 = 0.001$. (Software gives the more precise value $p \approx 0.0008$.)

(b) Since $p < 0.001$, the null hypothesis is rejected at $\alpha = 0.05$, at $\alpha = 0.01$, and at $\alpha = 0.001$. The result is "highly significant" by any conventional standard, and we can confidently conclude that the mean monthly food expenditure of urban households in Punjab differs from that of rural households.

Example 5.6(b) (p-Value for a Two-Sample Z-Test)

In Example 5.5 (the microwave-ownership study), we computed Z = 2.303 in a two-tailed test of H_0: $\pi_1 = \pi_2$. Compute the p-value of this test, and interpret it.

Solution. For a two-tailed test, the p-value is twice the area in one tail beyond the absolute value of the test statistic:

$$p = 2 \times P(Z \geq 2.303) = 2 \times (1 - 0.9893) = 2 \times 0.0107 = 0.0214.$$

The p-value is approximately 0.021. This means that, if there were really no difference in microwave ownership between the two cities, we would obtain a difference at least as large as the one we observed in only about 2.1% of all possible random samples of these sizes. The result is "moderately significant": clearly significant at $\alpha = 0.05$ (since $0.021 < 0.05$), but not significant at $\alpha = 0.01$ (since $0.021 > 0.01$). A reader who normally uses $\alpha = 0.05$ would conclude that the two cities really do differ, while a reader who insists on $\alpha = 0.01$ would withhold judgement until more evidence is available.

5.10 CHOOSING THE RIGHT TWO-SAMPLE TEST: A DECISION MAP

With four different two-sample tests for means and one for proportions, the student may be feeling slightly bewildered. The following decision map summarizes which test to use in any given situation. The student should refer to it whenever a two-sample problem is encountered, and should use it explicitly until the choice becomes second nature.

1. Are the two samples independent or paired? If paired (matched, before-and-after, repeated measures on the same units), use the paired-difference t-test of Section 5.7. Stop. If independent, proceed to step 2.
2. Is the parameter of interest a mean or a proportion? If a proportion, use the pooled-proportion Z-test of Section 5.8. Stop. If a mean, proceed to step 3.
3. Are both samples large ($n_1 \geq 30$ and $n_2 \geq 30$)? If yes, use the large-sample Z-test of Section 5.4 (with sample standard deviations replacing the unknown population values). Stop. If no, proceed to step 4.
4. Can the two population variances be assumed equal? Check this by inspecting the sample standard deviations: if the larger one is no more than about twice the smaller one, the equal-variances assumption is reasonable. If yes, use the pooled-variance t-test of Section 5.5. If no, use the Welch unpooled t-test of Section 5.6.

This four-step procedure covers nearly every two-sample situation that arises in introductory social-science research. There are a few situations it does not cover — for example, comparing two variances, or comparing more than two means at once — and we shall meet those in Unit 6.

5.11 COMMON PITFALLS IN TWO-SAMPLE TESTING

Two-sample tests are even more vulnerable to misuse than one-sample tests, because there are more opportunities to make wrong choices. The student should be aware of the following common errors and should take active steps to avoid them.

5.11.1 Treating Paired Data as Independent (or Vice Versa)

This is by far the commonest error in two-sample inference. A researcher who measures the blood pressure of 30 patients before and after a treatment, and then runs a two-independent-samples t-test on the data, is throwing away a great deal of information and is likely to find a much smaller test statistic — and a much weaker conclusion — than the data really warrant. Conversely, a researcher who treats two clearly independent samples (say, 30 men and 30 women) as if they were paired is technically incorrect and may obtain a misleadingly extreme result. The remedy is simple: think carefully about how the data were collected before choosing the test. If each observation in sample 1 is naturally linked to one specific observation in sample 2, the data are paired. Otherwise, they are independent.

5.11.2 Ignoring Unequal Variances

When the two sample standard deviations are very different, the pooled-variance t-test is no longer valid, and using it can give either an inflated Type I error rate (a "false alarm" that is more likely than the nominal α) or a deflated power. Inspect the two sample standard deviations whenever you contemplate a two-sample t-test, and if their ratio exceeds about 2 (or if a formal F-test rejects the equal-variances assumption), use the Welch unpooled procedure instead. The Welch test is only marginally less powerful when variances are equal but is much more reliable when they are not.

5.11.3 Comparing Samples from Different Populations

A two-sample test is valid only if the two samples have been drawn at random from their respective populations. A study that compares 40 volunteers from a wealthy neighborhood with 40 volunteers from a poor neighborhood is not really comparing two populations: it is comparing two convenience samples that may differ in many ways besides the variable of interest. The result of such a study cannot be generalized to the broader populations, and any "significant" finding may simply reflect uncontrolled differences between the two groups. The remedy is to use random sampling from clearly defined populations whenever possible, and to report the limitations of the design honestly when it is not.

5.11.4 Multiple Comparisons in Two-Sample Studies

When a researcher has several response variables and runs a separate two-sample test for each, the probability that at least one of them produces a "significant" result by chance alone grows quickly. A study that compares two groups on 20 different outcome variables, each tested at $\alpha = 0.05$, is almost guaranteed to find at least one "significant" difference even if the two populations are identical in every way. This is exactly the multiple-testing problem we met in Section 4.13. The remedy, as before, is to apply a correction such as the Bonferroni procedure (use α/k as the per-test level when k tests are being carried out), or to plan the study around a single primary outcome rather than fishing through many.

5.12 SUMMARY

This unit has developed the inferential techniques needed to compare two populations on the basis of two samples. The main points may be summarized as follows.

- Two samples are independent if the units in the two samples are completely separate; they are paired (dependent) if each observation in one sample is naturally linked to a specific observation in the other. The choice between independent-samples tests and paired tests must reflect how the data were actually collected.

- The sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ for independent samples has mean $(\mu_1 - \mu_2)$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. By the Central Limit Theorem, this sampling distribution is approximately normal for sufficiently large samples, regardless of the shape of the populations.
- The large-sample Z-test of Section 5.4 is used when both $n_1 \geq 30$ and $n_2 \geq 30$. The sample standard deviations may be substituted for the unknown population standard deviations.
- The pooled-variance t-test of Section 5.5 is used when both samples are small but the two populations can be assumed normal with equal variances. The pooled sample variance s_p^2 is the weighted average of s_1^2 and s_2^2, and the test statistic has $df = n_1 + n_2 - 2$ degrees of freedom.
- The Welch unpooled t-test of Section 5.6 is used when the two populations can be assumed normal but their variances are clearly unequal. The Welch–Satterthwaite formula gives an approximate (and usually fractional) number of degrees of freedom.
- The paired-difference t-test of Section 5.7 reduces a paired two-sample problem to a one-sample t-test on the within-pair differences. It has $df = n - 1$ (where n is the number of pairs) and is typically more powerful than the corresponding independent-samples test, because much of the between-individual variability is removed by pairing.
- For testing the difference between two population proportions in large samples, the pooled-proportion Z-test of Section 5.8 is used, with the pooled estimate $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$ appearing in the standard error formula. Notice that pooling is appropriate for hypothesis tests but not for confidence intervals; in the latter case, the unpooled formula of Section 3.10 should be used.
- The decision map of Section 5.10 helps the student choose the right test in any given situation: paired or independent? mean or proportion? large or small sample? equal or unequal variances?
- The most common pitfalls in two-sample testing are (a) treating paired data as independent or vice versa, (b) ignoring unequal variances when applying the pooled t-test, (c) comparing samples drawn from different populations rather than from the same population, and (d) ignoring the multiple-comparisons problem when many tests are run on the same data set.

In Unit 6 we shall extend the same general framework to chi-square tests (for goodness of fit and independence) and to the analysis of variance (for comparing more than two means simultaneously). The student who has mastered the material of Units 4 and 5 will find the transition to those more advanced techniques very natural.

5.13 SELF-ASSESSMENT EXERCISE

1. Distinguish clearly between independent samples and paired (dependent) samples. Give one example of each from social-science research.
2. What is the formula for the standard error of the difference between two independent sample means? Why do the variances add even though the means subtract?
3. State the three assumptions of the pooled-variance t-test. Which assumption is the most restrictive, and what alternative test should be used when it fails?
4. Why is the paired-difference t-test typically more powerful than the corresponding two-independent-samples test? Explain the role of within-pair variability in your answer.

5. When testing H_0: $\pi_1 = \pi_2$ for two independent samples, why do we pool the two sample proportions in the standard error, whereas in constructing a confidence interval for $(\pi_1 - \pi_2)$ we do not pool? Justify the difference.

6. Develop the theory of the pooled-variance t-test for the difference between two population means, beginning with the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$, explaining why the Central Limit Theorem is not enough to justify the use of the standard normal distribution for small samples, deriving the pooled sample variance, writing down the test statistic, and stating its sampling distribution under the null hypothesis. State the assumptions clearly and discuss what should be done when each assumption fails.

7. Explain the difference between independent and paired samples, with particular reference to the standard error of the resulting test statistic. Show, by means of a small numerical example of your own construction, that treating paired data as independent can lead to a much smaller test statistic and possibly to a wrong conclusion. Then describe the standard procedure for the paired-difference t-test, including the formulas for $\bar{d}$, s_d, the test statistic, and the degrees of freedom.

8. A researcher computes a two-sample t-statistic of $t = 1.85$ with $df = 28$, in a one-tailed test. Without performing any further calculations, decide whether the result would lead to rejection of H_0 at $\alpha = 0.10$, at $\alpha = 0.05$, and at $\alpha = 0.01$. Explain your reasoning. [Hint: $t_{0.10,\,28} = 1.313$; $t_{0.05,\,28} = 1.701$; $t_{0.01,\,28} = 2.467$.]

9. Two independent samples gave the following statistics: $n_1 = 25$, $\bar{x}_1 = 50$, $s_1 = 8$ and $n_2 = 25$, $\bar{x}_2 = 45$, $s_2 = 18$. (a) Should you use a pooled-variance t-test or a Welch unpooled t-test on these data? Justify your choice. (b) Without doing the actual computation, state which test would have a smaller standard error, and explain why.

10. A study reports a two-sample 95% confidence interval for $(\mu_1 - \mu_2)$ of $(-2.5\ ;\ +6.0)$. Without doing any further calculation, decide whether the corresponding two-tailed test of H_0: $\mu_1 = \mu_2$ at $\alpha = 0.05$ would reject the null hypothesis. Explain.

11. A study comparing two teaching methods reports a t-statistic of $t = 3.20$ with $df = 38$ in a two-tailed test. The point estimate of the difference is 1.4 marks. Discuss whether this result is "statistically significant" and whether it is "practically significant", and explain the difference between the two concepts.

12. A researcher has 40 patients and wants to compare two diet programs. Discuss the relative merits of (a) randomly assigning 20 patients to each program (independent-samples design) versus (b) putting each of the 40 patients on both programs in a randomly assigned order (paired/repeated-measures design). Which design would you expect to give the more powerful test, and why?

13. A random sample of $n_1 = 50$ male university students at AIOU gave a mean monthly pocket money of $\bar{x}_1 =$ Rs. 9,400 with a sample standard deviation of $s_1 =$ Rs. 1,800. An independent random sample of $n_2 = 60$ female students gave $\bar{x}_2 =$ Rs. 8,800 with $s_2 =$ Rs. 1,500. Test, at $\alpha = 0.05$, whether the mean pocket money differs between male and female students. Use the large-sample Z-test. [Answer: $Z \approx 1.875$; $|Z| < 1.96$; fail to reject H_0]

14. Two independent samples of small businesses in two cities gave the following monthly profits (in thousand rupees). Sample 1 (Lahore): $n_1 = 16$, $\bar{x}_1 = 85$, $s_1 = 12$. Sample 2 (Faisalabad): $n_2 = 14$, $\bar{x}_2 = 78$, $s_2 = 10$. Assuming the two populations are normal with equal variances, test at $\alpha = 0.05$ whether the mean profits differ. [Answer: pooled t-test; $s_p \approx 11.10$; $t \approx 1.726$; $df = 28$; $t_{0.025,\,28} = 2.048$; fail to reject]

15. A pharmaceutical company claims that its new pain-relief tablet acts faster than the standard tablet. In a clinical trial, $n_1 = 12$ patients given the new tablet had a mean time-to-relief of $\bar{x}_1 = 18.5$ minutes with $s_1 = 4.2$ minutes, while $n_2 = 15$ patients given the standard tablet had $\bar{x}_2 = 22.8$ minutes with $s_2 = 5.0$ minutes. Assuming equal variances, test at $\alpha = 0.01$ whether the new tablet acts faster. [Answer: pooled t-test; $s_p \approx 4.66$; $t \approx -2.39$; df = 25; $t_{0.01\ 25} = 2.485$; fail to reject]

16. Ten students were given a coaching course before taking a standardized test. The scores (out of 100) before and after the course were:

Student	1	2	3	4	5	6	7	8	9	10
Before	62	58	70	55	65	60	72	68	64	59
After	68	62	74	60	69	64	78	71	67	65

Test, at $\alpha = 0.05$, whether the coaching course improved the mean score. Use the paired-difference t-test. [Answer: $\bar{d} = 4.6$; $s_d \approx 1.51$; $t \approx 9.65$; df = 9; reject H_0 very strongly]

17. A market researcher wishes to compare the proportion of households in two cities that own a washing machine. In a random sample of 250 households in Karachi, 175 own a washing machine; in an independent random sample of 300 households in Multan, 180 own a washing machine. Test, at $\alpha = 0.05$, whether the two proportions differ. [Answer: $\hat{p}_1 = 0.70$, $\hat{p}_2 = 0.60$, $\hat{p}$_pool = 0.6455; SE $\approx$ 0.0411; $Z \approx 2.43$; $|Z| > 1.96$; reject H_0]

18. A researcher in Khyber Pakhtunkhwa is comparing the literacy rates of two districts. In a random sample of $n_1 = 500$ adults from District A, 285 are literate. In an independent random sample of $n_2 = 600$ adults from District B, 318 are literate. Test, at $\alpha = 0.01$, whether the literacy rate differs between the two districts. [Answer: $\hat{p}_1 = 0.570$, $\hat{p}_2 = 0.530$, $\hat{p}$_pool = 0.5482; SE $\approx$ 0.0301; $Z \approx 1.33$; $|Z| < 2.576$; fail to reject]

19. A car manufacturer wishes to determine whether two different fuel-injection systems give the same average fuel economy. The manufacturer tests $n_1 = 8$ cars equipped with system A and finds a mean of $\bar{x}_1 = 14.5$ km/l with $s_1 = 1.2$ km/l. The same manufacturer tests $n_2 = 10$ cars equipped with system B and finds $\bar{x}_2 = 13.6$ km/l with $s_2 = 0.9$ km/l. Assuming the two populations are normal with equal variances, test at $\alpha = 0.05$ whether the mean fuel economies differ. [Answer: pooled t-test; $s_p \approx 1.04$; $t \approx 1.83$; df = 16; $t_{0.025,\ 16} = 2.120$; fail to reject]

20. In a sociological study, the sample mean income of $n_1 = 100$ male respondents was $\bar{x}_1 =$ Rs. 65,000 with $s_1 =$ Rs. 18,000. The sample mean income of $n_2 = 120$ female respondents was $\bar{x}_2 =$ Rs. 51,000 with $s_2 =$ Rs. 14,000. Test, at $\alpha = 0.01$, whether the mean income of men is at least Rs. 10,000 higher than that of women. (This is a one-tailed test of H_0: $\mu_1 - \mu_2 = 10{,}000$ versus H_1: $\mu_1 - \mu_2 > 10{,}000$.) [Answer: large-sample Z; SE $\approx$ 2210; $Z \approx (14000 - 10000)/2210 \approx 1.81$; one-tailed critical value 2.33; fail to reject]

21. Eight subjects were tested twice on a memory task, once after 4 hours of sleep and again, on a different day, after 8 hours of sleep. The numbers of items correctly recalled were:

Subject	1	2	3	4	5	6	7	8
4 hours	22	19	25	18	21	20	24	17
8 hours	27	23	29	22	25	26	30	21

Test, at $\alpha = 0.05$, whether more sleep improves memory performance. [Answer: $\bar{d} = -4.5$; $s_d \approx 0.756$; $t \approx -16.84$; df = 7; reject very strongly]

22. Two random samples were drawn from two different populations. Sample 1: $n_1 = 30$, $\bar{x}_1 = 100$, $s_1 = 8$. Sample 2: $n_2 = 25$, $\bar{x}_2 = 95$, $s_2 = 22$. (a) Determine whether the pooled-variance t-test or the Welch t-test is more appropriate. (b) Carry out the test you chose at $\alpha = 0.05$ (two-tailed). [Answer: Welch ($s_2/s_1 = 2.75 > 2$); SE = $\sqrt{(64/30 + 484/25)} = \sqrt{21.49} \approx 4.64$; $t \approx 1.078$; df ≈ 27 (approx); fail to reject]

23. A government agency claims that the proportion of households in Pakistan with access to clean drinking water is higher in urban areas than in rural areas. A random sample of 500 urban households finds 425 with access; an independent random sample of 600 rural households finds 432 with access. Test, at $\alpha = 0.05$, whether the urban proportion is higher. [Answer: $\hat{p}_1 = 0.85$, $\hat{p}_2 = 0.72$, $\hat{p}_pool \approx 0.779$; SE ≈ 0.0252; $Z \approx 5.16$; one-tailed critical value 1.645; reject very strongly]

24. A factory manager believes that two assembly lines produce items with the same mean weight but with different variances. The first assembly line's sample of 20 items gave $\bar{x}_1 = 250$ g with $s_1 = 5$ g; the second line's independent sample of 20 items gave $\bar{x}_2 = 252$ g with $s_2 = 12$ g. Carry out the appropriate test at $\alpha = 0.05$ to decide whether the two means are equal. State clearly which test you are using and why. [Answer: Welch t-test (variance ratio 2.4); SE = $\sqrt{(25/20 + 144/20)} = \sqrt{8.45} \approx 2.907$; $t \approx -0.688$; df ≈ 25 (approx); fail to reject]

25. A researcher claims that students who study with background music score at least 5 marks lower on average than students who study in silence. Two independent samples gave: with-music group ($n_1 = 36$) $\bar{x}_1 = 64$, $s_1 = 12$; silence group ($n_2 = 40$) $\bar{x}_2 = 72$, $s_2 = 10$. Test the researcher's claim at $\alpha = 0.05$. (Hint: this is a one-tailed test of H_0: $\mu_1 - \mu_2 = -5$ versus H_1: $\mu_1 - \mu_2 < -5$.) [Answer: SE ≈ 2.523; $Z \approx (-8 - (-5))/2.523 \approx -1.189$; one-tailed critical value -1.645; fail to reject]

UNIT 6: CHI-SQUARE AND ANALYSIS OF VARIANCE

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Use the chi-square distribution to test whether an observed frequency distribution agrees with a theoretical model (the goodness-of-fit test) and to test whether two categorical variables are independent (the test of independence).
2. Carry out a chi-square test of homogeneity to compare several populations with respect to a single categorical variable, and explain the conceptual difference between this test and the test of independence.
3. Use the chi-square distribution to test a hypothesis about a single population variance, and use the F-distribution to test the equality of two population variances.
4. Carry out a one-way analysis of variance (ANOVA) to test whether several population means are equal, including the computation of the total, between-groups and within-groups sums of squares, the corresponding mean squares, and the F-ratio.
5. State the assumptions of each of the tests in this unit, recognize the situations in which each test is appropriate, and identify the most common pitfalls that arise when these tests are applied to social-science data.

6.1 INTRODUCTION

In Units 4 and 5 we developed the standard one-sample and two-sample tests for population means and population proportions. The tests we considered there had two important features in common. First, they all dealt with quantitative variables — variables that can be measured on a numerical scale, such as income, weight or reaction time — or with single yes/no attributes that could be summarized as proportions. Second, they all compared at most two populations or two groups at a time. The present unit removes both of these restrictions. We shall develop techniques that apply to categorical variables with several categories, and techniques that compare three or more populations simultaneously.

The unit is divided into two main parts, corresponding to two distinct families of tests. The first part (Sections 6.3 to 6.7) is devoted to tests based on the chi-square (χ^2) distribution. The chi-square distribution is the natural tool for analyzing categorical data, and it appears in three different but closely related testing situations: the goodness-of-fit test (which checks whether an observed frequency distribution conforms to a theoretical model), the test of independence (which checks whether two categorical variables are statistically independent), and the test of homogeneity (which checks whether several populations have the same distribution over a categorical variable). The same chi-square distribution also serves as the sampling distribution for inferences about a single population variance, and a closely related distribution — the F-distribution — does the same for the ratio of two population variances.

The second part of the unit (Sections 6.8 to 6.12) develops the analysis of variance, universally abbreviated as ANOVA. ANOVA is the natural extension of the two-sample t-test to the case of three or more groups. A psychologist who wants to compare four different therapy techniques, an agricultural scientist who wants to compare five different fertilizers, or an educational researcher who wants to compare the effectiveness of three different teaching methods cannot use a two-sample t-test, because there are too many groups. Running a separate two-sample test for every possible pair of groups is also unsatisfactory, because the multiple-comparisons problem we met in Unit 5 makes the overall Type I error rate balloon out of control. ANOVA solves both problems by carrying out a single overall test of the hypothesis that all the group means are equal, while keeping the Type I error rate at the chosen value of α.

The level of difficulty in this unit is comparable to that of Unit 5. The student is expected to be comfortable with the six-step hypothesis-testing procedure and with the p-value approach, and to be able to make conscious choices among different tests on the basis of the structure of the data. The exercises at the end of the unit, like those at the end of Unit 5, contain both routine numerical problems and conceptual numerical problems that demand interpretation and judgement rather than mere computation.

6.2 THE CHI-SQUARE DISTRIBUTION: A BRIEF REVIEW

We met the chi-square distribution briefly in Section 3.11 in connection with the construction of confidence intervals for a population variance. Because the chi-square distribution will play such a central role in the rest of this unit, it will be useful to review its main properties before we proceed.

The chi-square distribution is a probability distribution that is defined only for non-negative values, is right-skewed, and depends on a single parameter called the degrees of freedom (df). For very small df the distribution is highly skewed; as df increases the distribution becomes more nearly symmetric and, for very large df, approaches a normal distribution. The mean of a chi-square distribution with df degrees of freedom is exactly df itself, and its variance is 2*df.

The critical values of the chi-square distribution for various tail probabilities and various degrees of freedom are tabulated in standard chi-square tables, which can be found in any introductory statistics textbook. The

notation $\chi^2_{\alpha,df}$ denotes the value with df degrees of freedom that has area α to its right (in the upper tail). For example, $\chi^2_{0.05,10}$ = 18.307, which means that, for a chi-square random variable with 10 degrees of freedom, the probability of exceeding 18.307 is exactly 0.05. All the chi-square tests in this unit will use this kind of upper-tail critical value, because the test statistic is always non-negative and large values constitute evidence against the null hypothesis.

6.3 CHI-SQUARE AS A TEST OF GOODNESS OF FIT

The chi-square goodness-of-fit test is used to decide whether an observed frequency distribution is consistent with a theoretical model. The data consist of a sample of n observations classified into k mutually exclusive and exhaustive categories. The theoretical model specifies, for each category, the proportion of the population that should fall in that category if the model is correct. We then compare the observed frequencies in the sample with the frequencies that would be expected under the model, and we ask whether the discrepancies between observed and expected are larger than what could plausibly be attributed to sampling error alone.

6.3.1 The Test Statistic

Let $O_1, O_2, \ldots, O_k$ denote the observed frequencies in the k categories of the sample, and let $E_1, E_2, \ldots, E_k$ denote the corresponding expected frequencies under the theoretical model. The expected frequency in category i is computed by multiplying the total sample size n by the model probability for that category: $E_i = n \cdot p_i$. The chi-square test statistic is then defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (6.1)$$

Each term in the sum is non-negative; the term is zero when the observed frequency in that category equals the expected frequency, and grows larger as the discrepancy between O and E grows larger. A small total χ^2 indicates good agreement between the observed data and the theoretical model; a large total χ^2 indicates poor agreement and provides evidence against the model.

Under the null hypothesis that the model is correct, the test statistic (6.1) follows approximately a chi-square distribution. The number of degrees of freedom depends on the number of categories k and on the number of parameters of the model that have to be estimated from the same sample data. Specifically:

$$df = k - 1 - m \qquad (6.2)$$

where m is the number of parameters of the model estimated from the sample. When all the model probabilities p_i are completely specified in advance (no parameters need to be estimated), m = 0 and the degrees of freedom are simply k − 1. When the model contains, say, an unknown mean that has to be estimated from the sample, m = 1 and df = k − 2. The reason for subtracting m is that each estimated parameter "uses up" one degree of freedom, just as in the formulas for the sample variance and the sample standard deviation.

6.3.2 The Decision Rule

The chi-square goodness-of-fit test is always a one-tailed (upper-tailed) test, because only large values of χ^2 indicate disagreement with the model. We reject the null hypothesis at significance level α if:

$$\chi^2 > \chi^2_{\alpha,df} \qquad (6.3)$$

and otherwise fail to reject. The conventional rule of thumb for the validity of the chi-square approximation is that every expected frequency E_i should be at least 5; if some categories have very small expected frequencies, neighboring categories should be combined until this requirement is met.

Example 6.1 (Goodness-of-Fit Test — Equal Probabilities)

A six-sided die is rolled 120 times to test whether it is fair. The observed frequencies of the six faces are as follows:

Face	1	2	3	4	5	6	Total
Observed (O)	15	24	18	22	17	24	120

Test, at the 5% level of significance, whether the die is fair.

Solution.

Step 1. Hypotheses.

$$H_0: \textit{the die is fair } (\textit{each face has probability } 1/6).$$

$$H_1: \textit{the die is not fair}.$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: chi-square goodness-of-fit (6.1). Under H_0, each expected frequency is $E_i = n \cdot p_i = 120 \cdot (1/6) = 20$.

Step 4. Degrees of freedom: $df = k - 1 - m = 6 - 1 - 0 = 5$ (no parameters estimated). Critical value: $\chi^2_{0.05,5} = 11.070$. We reject H_0 if $\chi^2 > 11.070$.

Step 5. Computation:

Face	O	E	O − E	$(O - E)^2$	$(O - E)^2/E$
1	15	20	−5	25	1.250
2	24	20	4	16	0.800
3	18	20	−2	4	0.200
4	22	20	2	4	0.200
5	17	20	−3	9	0.450
6	24	20	4	16	0.800
Total	120	120	0	—	3.700

So the test statistic is $\chi^2 = 3.700$.

Step 6. Decision and conclusion. Since $\chi^2 = 3.700 < 11.070$, we fail to reject H_0 at the 5% level. There is not sufficient evidence to conclude that the die is biased; the observed differences from the expected counts are consistent with the kind of random variation we would expect from a fair die in 120 rolls.

Example 6.2 (Goodness-of-Fit Test — Unequal Probabilities)

According to Pakistan's Bureau of Statistics, the four major provinces account for the following hypothetical proportions of the country's working population: Punjab 53%, Sindh 23%, Khyber Pakhtunkhwa 18%, Balochistan 6%. A random sample of n = 500 employees of a multinational company in Islamabad is classified by province of origin, with the following results:

Province	Punjab	Sindh	KPK	Balochistan	Total
Observed (O)	240	130	105	25	500

Test, at the 5% level, whether the multinational company's workforce reflects the national distribution of working population.

Solution.

Step 1. Hypotheses:

H_0: *the company's workforce follows the national proportions* $(0.53, 0.23, 0.18, 0.06)$.

H_1: *at least one of the proportions differs from the national value.*

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: equation (6.1).

Step 4. df = 4 − 1 − 0 = 3. Critical value $\chi^2_{0.05,3} = 7.815$. Reject H_0 if $\chi^2 > 7.815$.

Step 5. Compute the expected frequencies and the chi-square statistic:

Province	O	p	E = np	$(O - E)^2/E$
Punjab	240	0.53	265	2.358
Sindh	130	0.23	115	1.957
KPK	105	0.18	90	2.500
Balochistan	25	0.06	30	0.833
Total	500	1.00	500	7.648

Step 6. Decision and conclusion. Since $\chi^2 = 7.648 < 7.815$, we fail to reject H_0 at the 5% level — but only just. (At the 10% level, where the critical value is $\chi^2_{0.10,3} = 6.251$, the test would be significant.) There is no strong evidence at the conventional 5% level that the company's workforce departs from the national distribution, although the result is on the edge of significance and a larger sample might tip the conclusion the other way.

6.4 CHI-SQUARE AS A TEST OF INDEPENDENCE

A second use of the chi-square distribution is to test whether two categorical variables are independent of one another. The data here come from a single random sample of n units, each of which is classified simultaneously according to two categorical variables. The cross-classification of the sample by the two variables produces a two-way table called a contingency table, in which the rows correspond to the categories of one variable, the columns to the categories of the other variable, and the cells contain the joint observed frequencies. The question we wish to answer is whether the row classification and the column classification are independent in the population from which the sample was drawn.

6.4.1 Setting Up the Problem

Suppose the row variable has r categories and the column variable has c categories, so that the contingency table is r × c (read as "r by c"). Let O_{ij} denote the observed frequency in the cell at row i and column j (for i = 1, …, r and j = 1, …, c). Let R_i denote the i-th row total (the sum of all the observed frequencies in row i), C_j the j-th column total, and n the grand total (the sum of all the observed frequencies in the table). The hypotheses for the test of independence are:

1. H_0: the two variables are independent in the population.
2. H_1: the two variables are not independent (they are associated).

Under H_0, the probability that an observation falls in cell (i, j) is the product of the marginal probabilities for row i and column j. Estimating these marginal probabilities by R_i/n and C_j/n respectively, the expected frequency in cell (i, j) under independence is:

$$E_{ij} = \frac{R_i \cdot C_j}{n} \qquad (6.4)$$

6.4.2 The Test Statistic and Degrees of Freedom

The chi-square test statistic for the test of independence has the same general form as for the goodness-of-fit test, but the sum is now taken over all r × c cells of the contingency table:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (6.5)$$

Under H_0, this statistic follows approximately a chi-square distribution with degrees of freedom:

$$df = (r - 1)(c - 1) \qquad (6.6)$$

A heuristic justification for this formula is the following. The contingency table contains r × c cells, but the cell counts are constrained by the requirement that the row totals add up to R_i, the column totals to C_j, and the grand total to n. After fixing the marginal totals, only (r − 1)(c − 1) of the cell counts can vary freely; the rest are determined by the marginal sums. Hence the number of degrees of freedom is (r − 1)(c − 1).

The decision rule is the same as before: reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$, otherwise fail to reject. The same expected-frequency rule of thumb applies: every E_{ij} should be at least 5 for the chi-square approximation to be valid; if some cells have very small expected frequencies, related rows or columns should be combined.

Example 6.3 (Test of Independence — 2 × 3 Table)

A market researcher in Rawalpindi wishes to investigate whether there is an association between gender and the preferred type of mobile phone brand among adult consumers. A random sample of 500 adults is classified as follows:

Gender	Brand A	Brand B	Brand C	Total
Male	120	90	60	270
Female	80	85	65	230
Total	200	175	125	500

Test, at the 5% level of significance, whether gender and brand preference are independent.

Solution.

Step 1. Hypotheses:

$$H_0: \textit{gender and brand preference are independent.}$$

$$H_1: \textit{gender and brand preference are not independent.}$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: equation (6.5).

Step 4. df = (2 − 1)(3 − 1) = 2. Critical value $\chi^2_{0.05,2} = 5.991$. Reject H_0 if $\chi^2 > 5.991$.

Step 5. Compute the expected frequencies using equation (6.4):

1. E(Male, A) = (270 × 200)/500 = 108.0
2. E(Male, B) = (270 × 175)/500 = 94.5
3. E(Male, C) = (270 × 125)/500 = 67.5
4. E(Female, A) = (230 × 200)/500 = 92.0
5. E(Female, B) = (230 × 175)/500 = 80.5
6. E(Female, C) = (230 × 125)/500 = 57.5

Now compute the chi-square statistic:

Cell	O	E	$(O - E)^2/E$
(M, A)	120	108.0	1.333
(M, B)	90	94.5	0.214
(M, C)	60	67.5	0.833
(F, A)	80	92.0	1.565
(F, B)	85	80.5	0.252
(F, C)	65	57.5	0.978
Total	500	500.0	5.175

Step 6. Decision and conclusion. Since $\chi^2 = 5.175 < 5.991$, we fail to reject H_0 at the 5% level. There is not sufficient evidence to conclude that gender and brand preference are associated; the differences observed in this sample could plausibly be attributed to random sampling variation alone.

6.5 CHI-SQUARE AS A TEST OF HOMOGENEITY

A third closely related use of the chi-square distribution is the test of homogeneity. Although the formulas for the test statistic and the degrees of freedom are identical to those of the test of independence, the conceptual setting is different and the student should make sure he or she understands the distinction.

In a test of independence, the data come from a single random sample of n units, each of which is then classified according to two different categorical variables. We are asking whether the two variables are independent of one another in the single population from which the sample was drawn.

In a test of homogeneity, by contrast, the data come from several independent random samples drawn from several different populations (one sample from each population), and each unit in each sample is classified according to a single categorical variable. We are asking whether the distribution of that variable is the same across all the populations — that is, whether the populations are homogeneous with respect to the variable of interest.

A simple way to remember the difference is the following: in the test of independence, the marginal totals of the rows and columns are both random; in the test of homogeneity, one set of marginals (typically the row totals) is fixed in advance by the researcher when planning the sample sizes for each population. The arithmetic, however, is identical: the expected frequencies are computed by formula (6.4), the test statistic is given by formula (6.5), and the degrees of freedom are $(r-1)(c-1)$ in both cases. The student should not be misled by the conceptual difference into thinking that the calculations are also different.

Example 6.4 (Test of Homogeneity)

A health survey is conducted in three districts of Punjab to compare the proportions of adults in three categories of body-mass index (BMI): under-weight, normal, and over-weight. Independent random samples are drawn from each district, with the following results:

District	Under-weight	Normal	Over-weight	Total
Lahore	40	120	90	250
Multan	35	140	75	250
Sialkot	25	155	70	250
Total	100	415	235	750

Test, at the 5% level of significance, whether the BMI distribution is homogeneous across the three districts.

Solution.

Step 1. Hypotheses:

$$H_0: \textit{the three districts have the same distribution of BMI.}$$

$$H_1: \textit{at least two districts have different distributions.}$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: same as for independence, equation (6.5).

Step 4. $df = (3-1)(3-1) = 4$. Critical value $\chi^2_{0.05,4} = 9.488$. Reject H_0 if $\chi^2 > 9.488$.

Step 5. Expected frequencies. By formula (6.4):

District	Under-weight (E)	Normal (E)	Over-weight (E)
Lahore	33.33	138.33	78.33
Multan	33.33	138.33	78.33

Sialkot	33.33	138.33	78.33

(Each row has total 250, and each column expected total equals the column observed total.) Now compute $(O - E)^2/E$ for every cell:

Cell	O	E	$(O - E)^2/E$
(Lah, U)	40	33.33	1.334
(Lah, N)	120	138.33	2.430
(Lah, O)	90	78.33	1.738
(Mul, U)	35	33.33	0.084
(Mul, N)	140	138.33	0.020
(Mul, O)	75	78.33	0.142
(Sia, U)	25	33.33	2.082
(Sia, N)	155	138.33	2.009
(Sia, O)	70	78.33	0.886
Total	750	750	10.725

Step 6. Decision and conclusion. Since $\chi^2 = 10.725 > 9.488$, we reject H_0 at the 5% level. There is sufficient evidence that the three districts differ in the distribution of BMI categories. Looking at the table of $(O - E)^2 / E$ values, the largest contributions to the chi-square come from the cells (Lah, N), (Sia, U) and (Sia, N), which suggests that Lahore has somewhat fewer than expected in the "normal" category while Sialkot has somewhat fewer "under-weight" and more "normal" individuals than expected. A more detailed follow-up analysis would be needed to characterise these patterns precisely.

6.6 INFERENCE ABOUT A SINGLE POPULATION VARIANCE

We saw in Section 3.11 that the sampling distribution of $(n - 1)s^2/\sigma^2$ for a random sample drawn from a normal population is the chi-square distribution with $(n - 1)$ degrees of freedom. This result, which we used there to construct a confidence interval for σ^2, also gives us a test statistic for testing a hypothesis about σ^2. The procedure follows the same six-step pattern that we have used throughout Units 4 and 5.

6.6.1 The Test Statistic

To test H_0: $\sigma^2 = \sigma_0^2$ against an appropriate alternative, the test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \qquad (6.7)$$

which under H_0 follows a chi-square distribution with df = n − 1 degrees of freedom. The test depends critically on the assumption that the population is normal; unlike tests for the mean, the chi-square test for the variance is not protected by the Central Limit Theorem and may be seriously misleading when applied to non-normal populations, even with large samples.

The form of the critical region depends on the alternative hypothesis:

Alternative	Critical Region
$H_1: \sigma^2 \neq \sigma_0^2$ (two-tailed)	$\chi^2 > \chi^2_{\frac{\alpha}{2},n-1}$ or $\chi^2 < \chi^2_{1-\frac{\alpha}{2},n-1}$
$H_1: \sigma^2 > \sigma_0^2$ (right-tailed)	$\chi^2 > \chi^2_{\alpha,n-1}$
$H_1: \sigma^2 < \sigma_0^2$ (left-tailed)	$\chi^2 < \chi^2_{1-\alpha,n-1}$

Example 6.5 (Test for a Population Variance)

A textile manufacturer in Faisalabad claims that the variance in the breaking strength of his cotton thread does not exceed 4 kg² (i.e., the standard deviation does not exceed 2 kg). A random sample of n = 25 threads is tested and gives a sample variance of $s^2 = 6.5$ kg². Assume that breaking strengths are normally distributed. Test, at the 5% level of significance, whether the manufacturer's claim is justified.

Solution.

Step 1. Hypotheses. The claim is $\sigma^2 \leq 4$, so we look for evidence that $\sigma^2 > 4$:

$$H_0: \sigma^2 = 4 \quad versus \quad H_1: \sigma^2 > 4 \quad (right-tailed\ test).$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: equation (6.7) with df = 24.

Step 4. Critical region. For df = 24 at $\alpha = 0.05$ (right-tailed), $\chi^2_{0.05,24} = 36.415$. Reject H_0 if $\chi^2 > 36.415$.

Step 5. Computation:

$$\chi^2 = \frac{(25-1)(6.5)}{4} = \frac{156}{4} = 39.0.$$

Step 6. Decision and conclusion. Since $\chi^2 = 39.0 > 36.415$, we reject H_0 at the 5% level. There is sufficient evidence that the variance in the breaking strength of the cotton thread exceeds 4 kg². The manufacturer's claim is not supported by the data, and the production process should be reviewed to bring variability under control.

6.7 INFERENCE ABOUT TWO POPULATION VARIANCES: THE F-DISTRIBUTION

We now consider the problem of testing whether two population variances are equal. This question arises naturally in two contexts. First, it is of interest in its own right when the researcher wants to compare the variability of two processes (for example, the dispersion of returns of two investment portfolios, or the variability of test scores of two classes taught by different methods). Second, it is needed as a preliminary check when deciding whether to use the pooled-variance t-test or the Welch unpooled t-test of Unit 5: a formal test of $H_0: \sigma_1^2 = \sigma_2^2$ gives an objective basis for that decision.

6.7.1 The F-Distribution

The sampling distribution of the ratio of two independent sample variances drawn from two normal populations with equal variances is a probability distribution called the F-distribution, named after Sir Ronald A. Fisher who first studied it. The F-distribution has two parameters, denoted ν_1 and ν_2 (the Greek letter nu), called the numerator degrees of freedom and the denominator degrees of freedom respectively. Like the chi-square distribution, the F-distribution is right-skewed and is defined only for non-negative values; it depends on both df parameters and is therefore tabulated separately for each combination.

A useful identity relates the upper-tail and lower-tail critical values of the F-distribution: $F_{1-\alpha,\,\nu_1,\,\nu_2} = 1 / F_{\alpha,\,\nu_2,\,\nu_1}$. This identity allows us to read lower-tail critical values from a one-sided F-table that contains only upper-tail values, and we shall use it in the worked example below.

6.7.2 The Test Statistic

Let s_1^2 and s_2^2 be the sample variances of two independent random samples of sizes n_1 and n_2 drawn from two normal populations with population variances σ_1^2 and σ_2^2 respectively. To test H_0: $\sigma_1^2 = \sigma_2^2$, the test statistic is the ratio of the two sample variances:

$$F = \frac{s_1^2}{s_2^2} \qquad (6.8)$$

which under H_0 follows an F-distribution with $\nu_1 = n_1 - 1$ numerator degrees of freedom and $\nu_2 = n_2 - 1$ denominator degrees of freedom. By convention, the larger sample variance is placed in the numerator and the smaller in the denominator, so that the computed F is always at least 1. (When this convention is followed, only the upper tail of the F-distribution is needed for two-tailed tests, with the upper-tail probability doubled to obtain the two-tailed p-value.)

The form of the critical region depends on the alternative:

Alternative	Critical Region (with $s_1^2 \geq s_2^2$)
H_1: $\sigma_1^2 \neq \sigma_2^2$ (two-tailed)	$F > F_{\alpha/2,\,\nu_1,\,\nu_2}$
H_1: $\sigma_1^2 > \sigma_2^2$ (right-tailed)	$F > F_{\alpha,\,\nu_1,\,\nu_2}$

Example 6.6 (F-Test for Two Variances)

A psychologist wishes to compare the variability of reaction times for two groups of subjects. A random sample of $n_1 = 13$ male subjects gave a sample standard deviation of $s_1 = 18.5$ ms, and an independent random sample of $n_2 = 16$ female subjects gave a sample standard deviation of $s_2 = 12.0$ ms. Assume that reaction times are approximately normally distributed in both populations. Test, at the 5% level of significance, whether the two population variances are equal.

Solution.

Step 1. Hypotheses (two-tailed):

$$H_0: {\sigma_1}^2 = {\sigma_2}^2 \;\; versus \;\; H_1: {\sigma_1}^2 \neq {\sigma_2}^2.$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: equation (6.8). The two sample variances are:

$$s_1^2 = (18.5)^2 = 342.25, \quad s_2^2 = (12.0)^2 = 144.00.$$

Place the larger variance in the numerator: $s_1^2 = 342.25$ in the numerator, $s_2^2 = 144.00$ in the denominator.

Step 4. Critical region. For a two-tailed test at $\alpha = 0.05$ with $\nu_1 = 12$ and $\nu_2 = 15$, we look up $F_{0.025,\,12,\,15} = 2.96$. We reject H_0 if $F > 2.96$.

Step 5. Computation:

$$F = \frac{342.25}{144.00} = 2.377.$$

Step 6. Decision and conclusion. Since $F = 2.377 < 2.96$, we fail to reject H_0 at the 5% level. There is not sufficient evidence to conclude that the two population variances differ. The observed sample variances are consistent with the hypothesis that the male and female populations have the same underlying variability in reaction times. As a practical matter, this result tells the psychologist that, if she now wishes to compare the two means, she may safely use the pooled-variance t-test of Section 5.5.

6.8 ANALYSIS OF VARIANCE: AN INTRODUCTION

The two-sample t-tests of Unit 5 allow us to compare two population means, but they cannot be used directly when more than two populations need to be compared. Suppose, for example, that an educational researcher wishes to compare four different teaching methods to see whether they produce different mean test scores. With four groups there are six possible pairwise comparisons (Method 1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4). The naive approach of running six separate two-sample t-tests, each at $\alpha = 0.05$, has two serious problems.

The first problem is the multiple-comparisons problem we met in Unit 4: the probability that at least one of the six tests produces a "significant" result by chance alone, even if all four populations have the same mean, is much higher than 0.05. With six independent tests at $\alpha = 0.05$, the probability of at least one false alarm is approximately $1 - (0.95)^6 \approx 0.265$ — that is, 26.5% rather than the nominal 5%. The overall (or "family-wise") Type I error rate has ballooned by more than a factor of five.

The second problem is conceptual rather than technical. Even if we could control the family-wise error rate by some kind of correction, running six pairwise tests gives us six separate decisions to reconcile, with no clear answer to the original question "are the four methods all equivalent, or are some of them different?". A single overall test that delivers a single yes-or-no answer to that question would be much more useful.

The analysis of variance, universally abbreviated as ANOVA and developed by Sir Ronald A. Fisher in the 1920s and 1930s, solves both problems. ANOVA performs a single test of the joint null hypothesis "all the population means are equal" against the alternative "at least two of the population means are different". The test controls the overall Type I error rate at exactly α, and it does so by an ingenious construction that compares the variability between the group means to the variability within the groups. We develop the construction in detail in the next two sections.

6.9 THE ONE-WAY ANOVA MODEL AND NOTATION

Suppose we have k independent random samples, drawn from k different populations whose means are μ_1, μ_2, …, μ_k. The population that the i-th sample comes from is also called the i-th treatment group, or simply

the i-th group. Let n_i denote the size of the i-th sample, so that the total sample size is $N = n_1 + n_2 + \ldots + n_k$. Let x_{ij} denote the j-th observation in the i-th group ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$). The data structure can be displayed as follows:

Group	Observations	Sample Mean	Sample Size
1	$x_{11}, x_{12}, \ldots, x_{1n}$	$\bar{x}_1$	n_1
2	$x_{21}, x_{22}, \ldots, x_{2n}$	$\bar{x}_2$	n_2
⋮	⋮	⋮	⋮
K	x_{k1}_, x_{k2} , . . . , x_{kn}	$\bar{x}_k$	n_k

We define two key sample means. The mean of the i-th group is $\bar{x}_i = (1/n_i)\ \Sigma_j\ x_{ij}$, the average of the observations within group i. The grand mean is $\bar{x} = (1/N)\ \Sigma_i\ \Sigma_j\ x_{ij}$, the average of all the observations across all groups. The grand mean is a weighted average of the k group means, with weights equal to the group sizes.

6.9.1 The Model Assumptions

The one-way ANOVA test rests on three assumptions, which are similar to (but slightly more demanding than) those of the pooled-variance t-test:

1. Independent random samples. The k samples are drawn independently of one another, and the observations within each sample are also independent.
2. Normality. The k populations are each normal (or at least approximately so). The procedure is reasonably robust to mild departures from normality, especially when the group sizes are similar and not too small.
3. Homogeneity of variances (homoscedasticity). The k populations all share a common variance σ^2. This is the most restrictive of the three assumptions, and the validity of the ordinary one-way ANOVA depends critically on it. When the variances are clearly unequal, alternative procedures (such as Welch's ANOVA) are available; we shall not discuss them in this introductory course.

6.9.2 The Hypotheses

The hypotheses for one-way ANOVA are:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k \quad versus \quad H_1: not\ all\ \mu_i\ are\ equal. \tag{6.9}$$

Notice carefully what the alternative hypothesis says, and what it does not say. H_1 says that at least one of the k means differs from the others, but it does not say which one (or which ones), nor does it say that they are all different. A rejection of H_0 in a one-way ANOVA is therefore only the beginning of the analysis: it tells us that the means are not all equal, but it does not tell us where the differences lie. To find that out, we need to follow up the ANOVA with so-called post-hoc multiple-comparison procedures (such as Tukey's HSD, the Bonferroni correction, or Scheffé's method), which are beyond the scope of this introductory course but which the student should know to consult before drawing detailed conclusions.

6.10 COMPUTING THE ONE-WAY ANOVA: SST, SSB, SSW AND THE F-RATIO

The fundamental insight of analysis of variance is that the total variability in the data can be decomposed into two additive parts: the variability between the group means and the variability within the groups. We

measure each kind of variability by a sum of squared deviations (a "sum of squares"), and we compare them by means of an F-ratio. The decomposition is the algebraic heart of ANOVA, and it is worth understanding clearly.

6.10.1 Total Sum of Squares (SST)

The total sum of squares measures the total variability of all the observations around the grand mean. It is defined as:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}\right)^2 \qquad (6.10)$$

and has total degrees of freedom equal to N − 1, because all N observations are involved and only the constraint that they have a single grand mean is imposed.

6.10.2 Sum of Squares Between Groups (SSB)

The between-groups sum of squares (also called the treatment sum of squares, SSTr) measures how much the group means differ from the grand mean. It is defined as:

$$SSB = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \qquad (6.11)$$

Each term in this sum represents the squared distance from a group mean to the grand mean, weighted by the size of the group. SSB has k − 1 degrees of freedom — one less than the number of groups, because the group means are constrained by the grand mean.

6.10.3 Sum of Squares Within Groups (SSW)

The within-groups sum of squares (also called the error sum of squares, SSE, or residual sum of squares) measures how much the individual observations differ from their own group mean. It is defined as:

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 \qquad (6.12)$$

and has N − k degrees of freedom, because each of the k groups contributes n_i observations constrained by one group mean, giving a total of $(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = N - k$.

6.10.4 The Fundamental Identity

A key algebraic identity, which the student is asked to verify in the exercise, is that the total sum of squares decomposes exactly as the sum of the between-groups and within-groups sums of squares:

$$SST = SSB + SSW \qquad (6.13)$$

and the corresponding decomposition of the degrees of freedom is:

$$(N - 1) = (k - 1) + (N - k) \qquad (6.14)$$

This decomposition is the conceptual foundation of analysis of variance. It says that the total variability of the data can be split into a part that is explained by differences between the group means (SSB) and a part that is left over inside the groups (SSW). If the group means are all equal, the between-groups part should be small relative to the within-groups part; if the group means are very different from one another, the between-groups part should be large relative to the within-groups part. The F-ratio formalises this comparison.

6.10.5 Mean Squares and the F-Ratio

Dividing each sum of squares by its degrees of freedom gives the corresponding mean square:

$$MSB = \frac{SSB}{k-1} \quad , \quad MSW = \frac{SSW}{N-k} \qquad (6.15)$$

The mean square within (MSW) is in fact the pooled estimate of the common variance σ^2 that we used in the pooled t-test of Unit 5, generalized to k groups. The mean square between (MSB) is also an estimate of σ^2 when H_0 is true (all the group means are equal), but it tends to be larger than σ^2 when H_0 is false. The test statistic is the ratio:

$$F = \frac{MSB}{MSW} \qquad (6.16)$$

which under H_0 follows an F-distribution with $\nu_1 = k - 1$ numerator degrees of freedom and $\nu_2 = N - k$ denominator degrees of freedom. The decision rule is to reject H_0 if $F > F_{\alpha,\, k-1,\, N-k}$, and otherwise to fail to reject. The ANOVA F-test is always one-tailed (upper-tailed), because under H_0 the F-ratio is centred near 1, and only large values of F constitute evidence against H_0.

6.11 THE ANOVA TABLE AND THE DECISION RULE

The results of a one-way ANOVA are conventionally summarized in a table called the ANOVA table, which displays the sources of variation, the corresponding sums of squares, the degrees of freedom, the mean squares, the F-ratio, and the p-value (or the critical value, depending on the convention being followed). The general form of the table is the following:

Source of Variation	Sum of Squares	df	Mean Square	F-ratio
Between groups (treatments)	SSB	$k - 1$	$MSB = SSB/(k-1)$	$F = MSB/MSW$
Within groups (error)	SSW	$N - k$	$MSW = SSW/(N-k)$	—
Total	SST	$N - 1$	—	—

The student should always present the results of an ANOVA in this standard form, both because it makes the calculations easy to check and because it is the format expected by reviewers of academic papers. We illustrate the entire procedure with a full worked example.

Example 6.7 (One-Way ANOVA — Three Teaching Methods)

An educational researcher wishes to compare the effectiveness of three different teaching methods in improving the test scores of secondary-school students in mathematics. Sixteen students are randomly assigned to the three methods (5 to Method A, 6 to Method B, 5 to Method C), and at the end of the semester each student is given the same standardized mathematics test (out of 100). The scores are as follows:

Method A	**78**	**82**	**75**	**80**	**85**	**—**
Method B	70	74	68	72	76	78
Method C	88	92	85	90	95	—

Test, at the 5% level of significance, whether the three teaching methods produce different mean scores.

Solution.

Step 1. Hypotheses:

$$H_0: \mu_A = \mu_B = \mu_C \quad versus \quad H_1: at\ least\ two\ means\ are\ different.$$

Step 2. $\alpha = 0.05$.

Step 3. Test statistic: one-way ANOVA F-ratio (6.16). The numerator df = k − 1 = 2 and denominator df = N − k = 16 − 3 = 13.

Step 4. Critical region. From the F-table, $F_{0.05,\,2,\,13} = 3.806$. Reject H_0 if $F > 3.806$.

Step 5. Computation. We first compute the group means and the grand mean.

Group A: sum = 78 + 82 + 75 + 80 + 85 = 400, $n_A = 5$, $\bar{x}_A = 80.0$.

Group B: sum = 70 + 74 + 68 + 72 + 76 + 78 = 438, $n_B = 6$, $\bar{x}_B = 73.0$.

Group C: sum = 88 + 92 + 85 + 90 + 95 = 450, $n_C = 5$, $\bar{x}_C = 90.0$.

Grand sum = 400 + 438 + 450 = 1288, total N = 5 + 6 + 5 = 16, grand mean $\bar{x} = 1288/16 = 80.5$.

SSB:

$$SSB = 5(80.0 - 80.5)^2 + 6(73.0 - 80.5)^2 + 5(90.0 - 80.5)^2$$

$$= 5(0.25) + 6(56.25) + 5(90.25) = 1.25 + 337.50 + 451.25 = 790.00.$$

SSW: Compute the sum of squared deviations within each group:

Group A: $(78-80)^2 + (82-80)^2 + (75-80)^2 + (80-80)^2 + (85-80)^2 = 4 + 4 + 25 + 0 + 25 = 58$.

Group B: $(70-73)^2 + (74-73)^2 + (68-73)^2 + (72-73)^2 + (76-73)^2 + (78-73)^2 = 9 + 1 + 25 + 1 + 9 + 25 = 70$.

Group C: $(88-90)^2 + (92-90)^2 + (85-90)^2 + (90-90)^2 + (95-90)^2 = 4 + 4 + 25 + 0 + 25 = 58$.

$$SSW = 58 + 70 + 58 = 186.$$

Mean squares:

$$MSB = \frac{790.00}{2} = 395.00, \quad MSW = \frac{186}{13} = 14.31.$$

F-ratio:

$$F = \frac{395.00}{14.31} = 27.60.$$

ANOVA Table:

Source	SS	df	MS	F
Between groups	790.00	2	395.00	27.60
Within groups	186.00	13	14.31	—
Total	976.00	15	—	—

Step 6. Decision and conclusion. Since $F = 27.60 \gg 3.806$, we reject H_0 at the 5% level — and indeed at the 1% level ($F_{0.01,\,2,\,13} = 6.701$) and well beyond. There is overwhelming evidence that the three teaching methods produce different mean scores. Looking at the group means ($\bar{x}_A = 80$, $\bar{x}_B = 73$, $\bar{x}_C = 90$), it appears that Method C produces the highest scores and Method B the lowest, but a formal post-hoc comparison procedure (such as Tukey's HSD) would be needed to identify exactly which pairs of means are significantly different. The student is invited to verify the additivity identity SST = SSB + SSW: indeed, 976 = 790 + 186, as required.

6.12 ASSUMPTIONS OF ANOVA AND WHAT TO DO WHEN THEY FAIL

We listed the three assumptions of one-way ANOVA at the start of Section 6.9: independent random samples, normal populations, and a common variance across all groups. In practice some of these assumptions are inevitably violated to a greater or lesser extent, and the student needs to know what to do in each case.

6.12.1 Violations of Independence

Independence is the most important of the three assumptions and the hardest to repair if it is violated. If the observations within a group are not independent (for example, if they are repeated measurements on the same subjects, or if they come from a cluster of related individuals such as children in the same classroom), then the ordinary one-way ANOVA is invalid and may give misleading results. The remedy is to use a different procedure altogether — repeated-measures ANOVA, mixed-effects models, or hierarchical (multi-level) models — depending on the nature of the dependence. These topics are well beyond the scope of this introductory course.

6.12.2 Violations of Normality

Mild departures from normality have only a modest effect on the validity of the F-test, especially when the group sizes are similar and not too small. The F-test is reasonably robust to symmetric non-normal distributions, but it can become unreliable when the populations are highly skewed or have heavy tails. When normality is doubtful, the standard remedy is to apply a variance-stabilizing transformation (such as the logarithmic transformation for right-skewed positive data, or the arcsine transformation for proportions) to the original observations before carrying out the ANOVA. An alternative is to use a non-parametric test such as the Kruskal–Wallis H-test, which we shall meet briefly in Unit 8.

6.12.3 Violations of Homogeneity of Variances

Inequality of variances across groups (heteroscedasticity) is the most common violation of the ANOVA assumptions in social-science research. The F-test is reasonably robust to mild heteroscedasticity, especially when the group sizes are equal. When the variances differ markedly and the group sizes are unequal, however, the F-test can become very unreliable: the actual Type I error rate can be much higher (or sometimes much lower) than the nominal α. There are two standard remedies. The first is a variance-stabilising transformation, as for non-normality. The second is to use Welch's ANOVA, which is the multi-group analogue of the Welch unpooled t-test of Unit 5 and does not require the equal-variances assumption. Welch's ANOVA is now the default in many modern statistical software packages, although it is not always covered in introductory textbooks.

6.13 COMMON PITFALLS IN CHI-SQUARE AND ANOVA

The chi-square tests and the ANOVA procedures of this unit are vulnerable to several common errors. The student should be aware of the following pitfalls and should take active steps to avoid them.

6.13.1 Small Expected Frequencies in Chi-Square

All chi-square tests rely on an approximation that is reliable only when the expected frequencies are not too small. The conventional rule of thumb is that every expected frequency should be at least 5; some textbooks insist on at least 10 for greater safety. When some cells have expected frequencies below this threshold, the chi-square approximation breaks down and the test may give a misleadingly small p-value. The remedy is to combine related rows or columns until every expected frequency meets the threshold, or to use Fisher's exact test instead of the chi-square approximation (especially in 2 × 2 tables).

6.13.2 Confusing Independence with Homogeneity

As we noted in Section 6.5, the chi-square tests of independence and homogeneity have identical formulas but different conceptual settings. The student must be clear about which design is being used: a single sample classified two ways (independence), or several independent samples classified one way (homogeneity). Although the arithmetic is the same and the decision is unchanged, the wording of the conclusion is different, and using the wrong wording can confuse the reader of the research report.

6.13.3 Treating ANOVA Rejection as Identification

A common misinterpretation of a significant ANOVA result is to conclude that "all the group means are different from one another". The alternative hypothesis of one-way ANOVA says only that at least two of the means differ, not that they are all different. After a significant F-test, follow-up post-hoc multiple-comparison procedures are needed to determine which specific pairs of means are significantly different. Reporting "ANOVA was significant; therefore, the four methods all differ from one another" is a serious overstatement of what the test actually says.

6.13.4 Ignoring Heterogeneity of Variances in ANOVA

The ordinary one-way ANOVA assumes that the k populations all share a common variance. When this assumption is clearly violated and the group sizes are unequal, the ordinary F-test can have a substantially inflated Type I error rate. The standard advice is to inspect the group sample standard deviations (or to perform a formal test such as Levene's test for equality of variances) before running the ANOVA, and to use Welch's ANOVA or a transformation if the variances are markedly different.

6.14 SUMMARY

This unit has developed two major families of inferential procedures: chi-square tests for categorical data and analysis of variance for comparing several means simultaneously. The main points may be summarized as follows.

- The chi-square distribution is a right-skewed distribution defined for non-negative values. It depends on a single parameter, the degrees of freedom, and is the natural sampling distribution for sums of squared standardized normal variables. As the degrees of freedom increase, the chi-square distribution becomes more nearly normal.
- The chi-square goodness-of-fit test compares an observed frequency distribution with a theoretical model. The test statistic is $\chi^2 = \Sigma\ (O_i - E_i)^2/E_i$, with degrees of freedom $k - 1 - m$, where k is the

number of categories and m is the number of model parameters estimated from the same data. The test is one-tailed: large values of χ^2 constitute evidence against the model.

- The chi-square test of independence checks whether two categorical variables, measured on the same sample, are statistically independent. The same test, with identical formulas, can be used as a test of homogeneity, in which case the data come from several independent samples and the question is whether the populations have the same distribution. In both cases the expected frequencies are computed from the marginal totals as $E_{ij} = R_i C_j / n$, and the degrees of freedom are $(r - 1)(c - 1)$.
- The chi-square test for a single population variance uses $\chi^2 = (n - 1)s^2/\sigma_0^2 \sim \chi^2(n - 1)$ under H_0. The procedure depends critically on the assumption of normality and is not protected by the Central Limit Theorem.
- The F-distribution arises as the sampling distribution of the ratio of two independent sample variances drawn from normal populations with equal variances. It depends on two parameters, the numerator and denominator degrees of freedom, and is the natural sampling distribution for tests of equality of two variances and for the F-test in ANOVA.
- The F-test for two population variances uses $F = s_1^2/s_2^2$ (with the larger sample variance in the numerator) and follows $F(n_1 - 1, n_2 - 1)$ under H_0. The test is sometimes used as a preliminary check before deciding between the pooled and the unpooled t-test of Section 5.5.
- Analysis of variance (ANOVA) extends the comparison of two means to $k \geq 3$ groups. The fundamental identity SST = SSB + SSW partitions the total variability into a "between-groups" part (which captures the differences among the group means) and a "within-groups" part (which captures the residual variability). The F-ratio MSB/MSW is the test statistic, with $k - 1$ numerator and $N - k$ denominator degrees of freedom. The ANOVA F-test is always one-tailed.
- A significant ANOVA tells us that not all the group means are equal but does not say which ones differ. Post-hoc multiple-comparison procedures (Tukey, Bonferroni, Scheffé) are needed to identify the specific pairs of groups responsible for the rejection.
- The most common pitfalls in this unit are (a) small expected frequencies in chi-square tests, (b) confusing the independence and homogeneity settings, (c) over-interpreting a significant ANOVA as identifying all pairs of differences, and (d) ignoring heteroscedasticity in ANOVA.

In Unit 7 we shall develop regression and correlation analysis, which generalise the ideas of this unit to the case of two (or more) quantitative variables and which use a similar partition-of-variability framework. In Unit 8 we shall meet a number of non-parametric methods that serve as alternatives to the chi-square and ANOVA procedures developed here. The student who has mastered the material of the present unit will find both of these later units a natural continuation of what he or she already knows.

6.15 SELF-ASSESSMENT EXERCISE

1. Briefly describe the chi-square distribution and list its main properties. Why is the chi-square test always a one-tailed (upper-tailed) test?
2. Distinguish between the chi-square test of independence and the chi-square test of homogeneity. How does the underlying sampling design differ in the two cases? Why are the formulas identical?
3. What is the rule of thumb concerning the expected frequencies in a chi-square test? What should you do if the rule is violated in a particular cell of a contingency table?

4. State the three assumptions of one-way analysis of variance. Which of the three is the most restrictive, and what alternative procedures are available when it fails?

5. After a one-way ANOVA rejects the null hypothesis, why is it not correct to conclude that all pairs of group means are significantly different from one another? What additional analysis is needed to identify the specific pairs that differ?

6. Explain in detail the chi-square test of independence for a contingency table. State the hypotheses, derive the formula for the expected frequencies under independence, write down the test statistic and its sampling distribution, and explain how the degrees of freedom $(r - 1)(c - 1)$ are obtained. Illustrate the entire procedure with a complete numerical example of your own choosing.

7. Develop the theory of one-way analysis of variance. State the model, the assumptions, and the hypotheses; define SST, SSB, SSW and verify the fundamental identity SST = SSB + SSW; explain the meaning of the corresponding mean squares MSB and MSW; and write down the F-ratio and its sampling distribution under H_0. Discuss why MSB tends to be larger than MSW when the null hypothesis is false but is approximately equal to it when the null hypothesis is true.

8. A chi-square statistic of $\chi^2 = 8.40$ is computed from a 3×3 contingency table. Without doing any further calculation, decide whether the result would lead to rejection of H_0 at $\alpha = 0.10$, at $\alpha = 0.05$, and at $\alpha = 0.01$. Justify your answer using the appropriate critical values. [Hint: $\chi^2_{0.10,4}\} = 7.779$; $\chi^2_{0.05,4} = 9.488$; $\chi^2_{0.01,4} = 13.277$.]

9. In a one-way ANOVA with $k = 4$ groups and a total sample size of $N = 28$, the researcher reports SSB = 240 and SSW = 360. (a) Construct the complete ANOVA table. (b) Compute the F-ratio and decide at $\alpha = 0.05$ whether the four group means are significantly different. (c) Without doing any further computation, what does the F-ratio tell you about the relative size of the between-groups and within-groups variability? [Critical value: $F_{0.05,\,3,\,24} = 3.009$.]

10. A study reports a chi-square test of independence in a 2×4 contingency table with $\chi^2 = 9.50$. (a) State the degrees of freedom. (b) Without doing the actual calculation, decide whether the result is significant at $\alpha = 0.05$. [Hint: df = 3; $\chi^2_{0.05,3} = 7.815$.]

11. A researcher computes the F-ratio for a one-way ANOVA and finds $F = 4.20$ with 3 and 30 degrees of freedom. The total sample size is $N = 34$ across $k = 4$ groups. (a) State the conclusion at $\alpha = 0.05$. (b) The point estimate of the largest pairwise difference between two group means is 8 marks (out of 100). Discuss whether this result is "statistically significant" and whether it is "practically important", and explain the difference between the two concepts. [Critical value: $F_{0.05,\,3,\,30} = 2.922$.]

12. In a two-way contingency table, you obtain $\chi^2 = 14.0$ with df = 6. (a) Which is larger, the chi-square statistic or its expected value under H_0? Explain. (b) Compute the approximate p-value bracketed by chi-square table values at $\alpha = 0.05$, 0.025 and 0.01. [Hint: under H_0, $E(\chi^2) = df = 6$; $\chi^2_{0.05,6} = 12.59$; $\chi^2_{0.025,6} = 14.45$; $\chi^2_{0.01,6} = 16.81$.]

13. A bag of jelly beans is supposed to contain four colours in equal proportions: red, green, yellow and orange. A random sample of 80 jelly beans is drawn, and the observed counts are: red 25, green 15, yellow 22, orange 18. Test, at $\alpha = 0.05$, whether the four colours are present in equal proportions. [Answer: $E_i = 20$ each; $\chi^2 = 1.25 + 1.25 + 0.20 + 0.20 = 2.90$; df = 3; critical value 7.815; fail to reject]

14. A genetics experiment predicts that pea plants should produce flowers in three colours in the ratio 9 : 3 : 4 (purple : pink : white). In a random sample of 320 plants, the observed counts are: purple 175,

pink 65, white 80. Test, at $\alpha = 0.05$, whether the data are consistent with the predicted ratio. [Answer: E = 180, 60, 80; $\chi^2 \approx 0.139 + 0.417 + 0 \approx 0.556$; df = 2; critical 5.991; fail to reject]

15. A market researcher in Karachi cross-classifies 400 randomly selected adult shoppers by gender and by preferred mode of payment (cash, credit/debit card, mobile wallet). The contingency table is:

Gender	Cash	Card	Mobile Wallet	Total
Male	90	70	50	210
Female	110	50	30	190
Total	200	120	80	400

Test, at $\alpha = 0.05$, whether gender and preferred payment mode are independent. [Answer: $E_{M,Cash} = 105$, $E_{M,Card} = 63$, $E_{M,MW} = 42$; etc. $\chi^2 \approx 2.143 + 0.778 + 1.524 + 2.368 + 0.860 + 1.684 \approx 9.36$; df = 2; critical 5.991; reject]

16. In a study of educational attainment in three districts of Sindh, independent random samples of 200 adults from each district are classified as illiterate, primary, or secondary-or-higher:

District	Illiterate	Primary	Secondary+	Total
District 1	60	80	60	200
District 2	40	90	70	200
District 3	50	70	80	200
Total	150	240	210	600

Test, at $\alpha = 0.05$, whether the distribution of educational attainment is homogeneous across the three districts. [Answer: $E_{ij} = (\text{row} \cdot \text{col})/600$; e.g. $E_{1,Illit} = 50$; $\chi^2 \approx 8.25$ (approx); df = 4; critical 9.488; fail to reject]

17. A quality-control engineer at a tyre factory in Sialkot claims that the variance in the lifetimes of his tyres is at most 4,000 km². A random sample of n = 16 tyres is tested and gives a sample variance of $s^2 = 6{,}200$ km². Assuming tyre lifetimes are normally distributed, test the engineer's claim at $\alpha = 0.05$. [Answer: $\chi^2 = (15)(6200)/4000 = 23.25$; right-tailed critical $\chi^2_{0.05,15} = 24.996$; fail to reject]

18. Two independent samples were drawn from two normal populations to compare their variances. Sample 1: $n_1 = 21$, $s_1^2 = 144$. Sample 2: $n_2 = 25$, $s_2^2 = 64$. Test, at $\alpha = 0.05$ (two-tailed), whether the two population variances are equal. [Answer: $F = 144/64 = 2.25$; $\nu_1 = 20$, $\nu_2 = 24$; critical $F_{0.025,\,20,\,24} \approx 2.33$; fail to reject]

19. The yields (in kg per plot) of three varieties of wheat under identical growing conditions in a research field at Faisalabad are:

Variety A	45	42	48	44	46	—
Variety B	52	54	50	55	53	51
Variety C	48	46	49	47	45	—

Test, at $\alpha = 0.05$, whether the three varieties produce different mean yields. Use the one-way ANOVA F-test. [Answer: $\bar{x}_A = 45$, $\bar{x}_B = 52.5$, $\bar{x}_C = 47$, grand mean ≈ 48.25; SSB ≈ 165.45; SSW ≈ 38.50; $F \approx (165.45/2)/(38.50/13) \approx 27.9$; reject very strongly]

20. A social scientist tests four different methods of presenting a survey questionnaire on response rate. Each method is tested on five randomly selected groups of households, and the response rates (in per cent) are:

Method 1	62	58	65	60	64
Method 2	70	74	68	72	66
Method 3	58	55	62	60	57
Method 4	66	70	68	64	62

Test, at $\alpha = 0.05$, whether the four methods produce different mean response rates. [Answer: $\bar{x}_1 = 61.8$, $\bar{x}_2 = 70.0$, $\bar{x}_3 = 58.4$, $\bar{x}_4 = 66.0$; grand mean = 64.05; SSB ≈ 384.55; SSW ≈ 144.40; $F \approx (384.55/3)/(144.40/16) \approx 14.20$; critical $F_{0.05,\,3,\,16} = 3.239$; reject]

21. A consumer-protection group tests the proportion of defective products in three different brands of light bulbs. Independent random samples are drawn:

Brand	Sample size	Defective
A	200	14
B	250	24
C	300	21

Test, at $\alpha = 0.05$, whether the proportion of defective bulbs is the same across the three brands. (Hint: this is a chi-square test of homogeneity on a 3 × 2 contingency table.) [Answer: row totals 200, 250, 300; col totals 59, 691; $E_{A,D} = 200 \cdot 59/750 \approx 15.73$, etc.; $\chi^2 \approx 1.71$ (approx); df = 2; critical 5.991; fail to reject]

22. A random sample of n = 30 measurements from a normal population gives a sample variance of $s^2 = 18.5$. Test, at $\alpha = 0.05$, whether the population variance differs from $\sigma_0^2 = 12$. (Two-tailed test.) [Answer: $\chi^2 = (29)(18.5)/12 = 44.71$; two-tailed critical values $\chi^2_{0.025,29} = 45.722$ and $\chi^2_{0.975,29} = 16.047$; fail to reject — barely]

23. In a one-way ANOVA with k = 3 groups, $n_1 = n_2 = n_3 = 8$, the sample means are $\bar{x}_1 = 50$, $\bar{x}_2 = 56$, $\bar{x}_3 = 60$, and the within-group sum of squares is SSW = 252. (a) Compute the grand mean and SSB. (b) Construct the ANOVA table. (c) Test at $\alpha = 0.05$ whether the three means are equal. [Answer: grand mean = 55.33; $SSB = 8[(50-55.33)^2 + (56-55.33)^2 + (60-55.33)^2] = 8(28.41 + 0.45 + 21.81) = 405.36$; $df_B = 2$, $df_W = 21$; MSB = 202.68, MSW = 12.0; F = 16.89; critical 3.467; reject]

24. Two independent samples are drawn to compare the variances of household monthly expenditure in two cities. From Lahore: $n_1 = 16$, $s_1^2 = 25{,}000{,}000$. From Karachi: $n_2 = 13$, $s_2^2 = 9{,}000{,}000$. Test, at $\alpha = 0.10$ (two-tailed), whether the population variances differ. [Answer: $F = 25/9 \approx 2.778$; $\nu_1 = 15$, $\nu_2 = 12$; critical $F_{0.05,\,15,\,12} \approx 2.62$; reject]

25. In a goodness-of-fit test for a Poisson distribution, the sample mean is used to estimate the unknown rate parameter λ. If the data are grouped into k = 6 categories, what are the appropriate degrees of freedom for the chi-square test? [Answer: df = k − 1 − m = 6 − 1 − 1 = 4]

26. A school principal wants to compare the test scores of four classes taught by four different teachers. The four sample sizes are $n_1 = n_2 = n_3 = n_4 = 10$, and the four sample standard deviations are $s_1 = 8$, $s_2 = 9$, $s_3 = 7$, $s_4 = 18$. Should the principal use the ordinary one-way ANOVA on these data, or is some other procedure preferable? Justify your answer in one or two sentences. (Note: this is a conceptual question — no F-statistic computation is required.) [Answer: the largest s (18) is more than twice the smallest (7), so the equal-variances assumption is violated. Use Welch's ANOVA (or apply a variance-stabilising transformation), not the ordinary ANOVA.]

UNIT 7: REGRESSION AND CORRELATION ANALYSIS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Construct a scatter diagram for bivariate data and identify whether the relationship between the two variables appears to be linear, non-linear, or non-existent.
2. Estimate the parameters of a simple linear regression equation by the method of least squares, interpret the slope and the intercept, and use the fitted line to predict the value of the dependent variable for a given value of the independent variable.
3. Compute and interpret the standard error of estimate, the coefficient of determination (R^2) and the Pearson product-moment correlation coefficient (r), and explain the mathematical relationship between R^2 and r.
4. Carry out hypothesis tests and construct confidence intervals for the population regression slope and the population correlation coefficient, and distinguish between statistical significance and practical usefulness of a regression relationship.
5. Describe the basic ideas of multiple regression analysis, including the estimation of the multiple regression equation, the interpretation of partial regression coefficients, the multiple coefficient of determination R^2, and the overall F-test.

7.1 INTRODUCTION

In all the inferential procedures we have studied so far — in Units 3 through 6 — the variables of interest have been treated one at a time. We have estimated the mean of a single variable, tested whether two population means differ, checked whether two categorical variables are independent, and compared the means of several groups. In real-world research, however, one of the most important questions a social scientist can ask is not about a single variable in isolation but about the relationship between two (or more) variables. Does household expenditure on food increase as income increases? Is the crime rate in a city related to the unemployment rate? Do students who study more hours score higher marks on an examination? In each of these questions the researcher is interested in whether — and how — one variable changes as another variable changes.

The present unit develops the two principal statistical techniques for studying such relationships: regression analysis and correlation analysis. Although the two are closely related and are almost always discussed together, they answer slightly different questions. Regression analysis models the functional relationship between the variables and allows us to predict the value of one variable (the dependent variable) from the value of another (the independent variable). Correlation analysis measures the strength and direction of the linear association between the two variables, without specifying which one is the "cause" and which is the "effect". Together, the two techniques form one of the most powerful and widely used tools in the entire social-science research toolkit.

The unit is organized in two main parts. The first part (Sections 7.2 to 7.11) develops simple linear regression and correlation — the case in which there is only one independent variable. The second part (Sections 7.12 to 7.15) extends the ideas to multiple regression and correlation, in which two or more independent variables are used simultaneously to predict the dependent variable. The level of difficulty is comparable to that of Unit 6. The student is expected to be comfortable with the idea of a linear equation, with the summation notation, and with the six-step hypothesis-testing procedure.

7.2 SCATTER DIAGRAMS AND THE IDEA OF A LINEAR RELATIONSHIP

The natural starting point for any study of the relationship between two quantitative variables is a scatter diagram (also called a scatter plot). A scatter diagram is simply a graph in which each observation is plotted as a single point, with the horizontal axis representing the value of one variable (called the independent variable, or the explanatory variable, conventionally denoted by x) and the vertical axis representing the value of the other variable (called the dependent variable, or the response variable, conventionally denoted by y). The resulting cloud of points gives a visual picture of the nature of the relationship: its direction (positive or negative), its form (linear, curved, or nonexistent), and its strength (how tightly the points cluster around the overall pattern).

Three distinct patterns are commonly observed:

1. A positive linear relationship: as x increases, y also tends to increase, and the points cluster around a line that slopes upward from left to right. Example: as household income increases, expenditure on education also tends to increase.
2. A negative linear relationship: as x increases, y tends to decrease, and the points cluster around a line that slopes downward from left to right. Example: as the price of a product increases, the quantity demanded tends to decrease.
3. No apparent relationship: the points form a shapeless cloud with no clear pattern, indicating that knowing x tells us nothing useful about y.

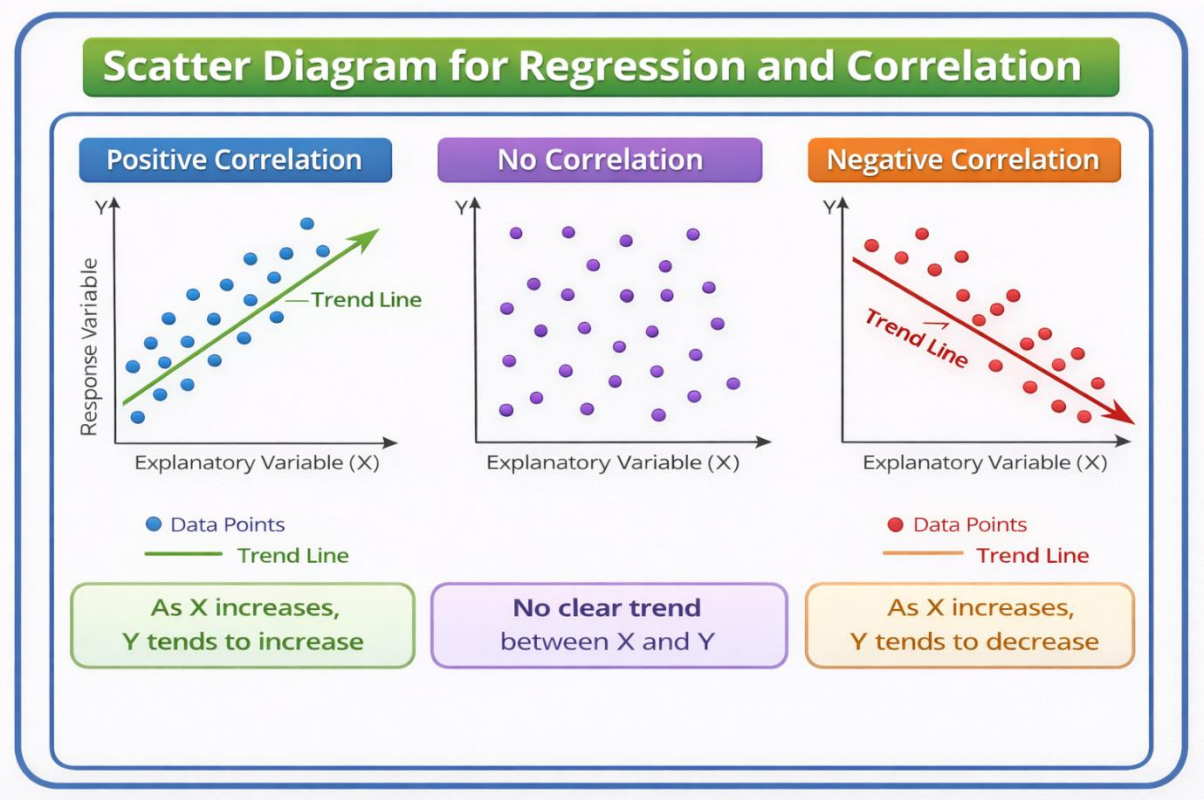


Figure 7.1 Scatter diagram: visual guide to relationship between variables

The scatter diagram is an indispensable first step in any regression or correlation analysis, because it reveals at a glance whether the assumptions of linearity that underlie the techniques of this unit are reasonable for the data at hand. A regression line fitted to data that are clearly non-linear will give misleading results.

7.3 THE SIMPLE LINEAR REGRESSION MODEL

The simple linear regression model describes the relationship between a dependent variable y and a single independent variable x by means of a straight line. In the population, the model is written as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i = 1, 2, \dots, n \qquad (7.1)$$

where:

1. Y_i is the value of the dependent variable for the i-th observation, treated as a random variable.
2. x_i is the value of the independent variable for the i-th observation, treated as known (non-random).
3. β_0 (beta-zero) is the population intercept — the predicted value of Y when x = 0. It is an unknown population parameter.
4. β_1 (beta-one) is the population slope — the change in the predicted value of Y for each one-unit increase in x. It is the single most important quantity in simple regression.
5. ε_i (epsilon) is the random error term. The standard assumptions are: $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ (constant), ε_i are independent and normally distributed.

7.3.1 The Assumptions (LINE)

The four assumptions — Linearity, Independence, Normality, Equal variance (LINE) — must hold for the inferential procedures of Sections 7.9 and 7.10 to be valid. When all four hold, the ordinary least-squares (OLS) estimates have the strongest optimality properties: they are unbiased, they are the most efficient

among all unbiased linear estimators (the Gauss–Markov theorem), and they allow the construction of exact t-tests and F-tests for the regression parameters.

7.4 ESTIMATION OF THE REGRESSION COEFFICIENTS BY THE METHOD OF LEAST SQUARES

The population regression line $E(Y) = \beta_0 + \beta_1 x$ is unknown, because the parameters β_0 and β_1 are unknown. We estimate them from the sample of n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ by the method of least squares, which chooses the values of the intercept and slope that minimize the sum of the squared vertical deviations of the observed y-values from the fitted line. The resulting estimated regression line, or sample regression line, is written:

$$\hat{y} = b_0 + b_1 x \qquad (7.2)$$

where b_0 and b_1 are the sample estimates of β_0 and β_1 respectively, and $\hat{y}$ (read as "y-hat") is the predicted value of y for a given x. The method of least squares minimizes:

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \qquad (7.3)$$

Taking partial derivatives of S with respect to b_0 and b_1 and setting them equal to zero yields the so-called normal equations. Solving these simultaneously gives the least-squares estimates:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad (7.4)$$

$$b_0 = \bar{y} - b_1 \bar{x} \qquad (7.5)$$

where $\bar{x} = \Sigma x_i/n$ and $\bar{y} = \Sigma y_i/n$ are the sample means of x and y. Equation (7.4) is the most important computational formula in this unit; the student should memorize it or at least know how to look it up quickly. Equation (7.5) shows that the estimated regression line always passes through the point $(\bar{x}, \bar{y})$, which is the centre of gravity of the scatter plot.

An equivalent and sometimes more convenient form of the slope formula uses the sums of squares and cross-products:

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \quad where \quad SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \;,\; SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (7.6)$$

Example 7.1 (Fitting a Simple Linear Regression)

A researcher at Allama Iqbal Open University wishes to study the relationship between the number of hours studied per week (x) and the marks obtained in an end-of-semester examination (y, out of 100) for a sample of n = 8 students of BS Economics. The data are as follows:

Student	1	2	3	4	5	6	7	8
Hours (x)	10	15	12	20	8	18	14	16
Marks (y)	52	68	58	80	45	75	62	70

Solution. We first prepare the calculation table.

x	Y	xy	x^2	y^2
10	52	520	100	2704
15	68	1020	225	4624
12	58	696	144	3364
20	80	1600	400	6400
8	45	360	64	2025
18	75	1350	324	5625
14	62	868	196	3844
16	70	1120	256	4900
$\Sigma = 113$	$\Sigma = 510$	$\Sigma = 7534$	$\Sigma = 1709$	$\Sigma = 33486$

The sample means are $\bar{x} = 113/8 = 14.125$ and $\bar{y} = 510/8 = 63.75$. Using equation (7.4):

$$b_1 = \frac{8(7534) - (113)(510)}{8(1709) - (113)^2} = \frac{60272 - 57630}{13672 - 12769} = \frac{2642}{903} = 2.926.$$

Using equation (7.5):

$$b_0 = 63.75 - 2.926(14.125) = 63.75 - 41.33 = 22.42.$$

The estimated regression equation is therefore:

$$\hat{y} = 22.42 + 2.926\,x.$$

Interpretation. The slope $b_1 = 2.926$ means that, for each additional hour studied per week, the predicted examination marks increase by approximately 2.93 marks. The intercept $b_0 = 22.42$ is the predicted marks for a student who studies zero hours per week; while this value is not practically meaningful (no real student studies zero hours), it anchors the left end of the regression line. For example, a student who studies 17 hours per week is predicted to score $\hat{y} = 22.42 + 2.926(17) = 22.42 + 49.74 \approx 72.2$ marks.

7.5 INTERPRETING THE REGRESSION EQUATION: SLOPE AND INTERCEPT

The slope b_1 of the sample regression line has a very clear and important interpretation: it is the estimated change in the dependent variable y for each one-unit increase in the independent variable x. If b_1 is positive, the relationship is positive (y increases as x increases); if b_1 is negative, the relationship is negative (y decreases as x increases); if b_1 is zero, there is no linear relationship at all. The magnitude of b_1 tells us the rate of change. In Example 7.1, each additional hour of study is associated with an increase of about 2.93 marks — a rate of change that is both positive and of a useful size.

The intercept b_0 is the predicted value of y when $x = 0$. It is sometimes meaningful (for example, the predicted starting salary of a new graduate with zero years of experience) and sometimes not (for example, the predicted marks for a student who does zero hours of study, as in Example 7.1). When $x = 0$ lies outside the range of the observed data, the intercept should be treated as a mathematical artefact that anchors the line

rather than as a meaningful substantive finding. This is a specific instance of the general rule against extrapolation, which we shall discuss further in Section 7.11.

7.6 THE STANDARD ERROR OF ESTIMATE

The standard error of estimate, denoted by s_e (or sometimes $s_{y|x}$), measures the average size of the residuals — the vertical distances between the observed y-values and the values predicted by the regression line. It is the regression analogue of the sample standard deviation: just as s measures the dispersion of a single variable around its mean, s_e measures the dispersion of the observed y-values around the regression line. The formula is:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \qquad (7.7)$$

The divisor is (n − 2) because two degrees of freedom have been "spent" in estimating the two parameters b_0 and b_1. The quantity in the numerator, $\Sigma(y_i - \hat{y}_i)^2$, is called the residual sum of squares (SSE). A computing formula that avoids the need to compute every predicted value individually is:

$$SSE = SS_{yy} - b_1 SS_{xy} \quad where \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (7.8)$$

Example 7.2 (Standard Error of Estimate)

Compute the standard error of estimate for the data of Example 7.1.

Solution. We first compute SSyy and SSxy.

$$SS_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 33486 - (510)^2/8 = 33486 - 32512.5 = 973.5.$$

$$SS_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n = 7534 - (113)(510)/8 = 7534 - 7203.75 = 330.25.$$

Using equation (7.8):

$$SSE = 973.5 - 2.926(330.25) = 973.5 - 966.11 = 7.39.$$

$$s_e = \sqrt{\frac{7.39}{8-2}} = \sqrt{1.232} = 1.110\ marks.$$

The standard error of estimate is approximately 1.11 marks. This means that the observed marks typically differ from the regression-predicted marks by about 1.11 marks, which is quite small relative to the range of the y-values (45 to 80). The regression line fits the data very well.

7.7 THE COEFFICIENT OF DETERMINATION (R^2)

The coefficient of determination, denoted by R^2 (or r^2 in simple regression), is the single most popular summary measure of how well a regression line fits the data. It is defined as the proportion of the total variation in y that is explained by the regression on x:

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SSR}{SS_{yy}} \qquad (7.9)$$

where SSR = SSyy − SSE is the regression sum of squares — the part of the total variation that is "explained" by the regression line. R^2 always lies between 0 and 1. A value near 1 means that the regression line explains

nearly all the variation in y; a value near 0 means that the regression explains very little. In the social sciences, R^2 values in the range 0.20 to 0.60 are common and perfectly acceptable; values above 0.90 are rare and sometimes indicate an over-fitted model or a near-tautological relationship.

Example 7.3 (Coefficient of Determination)

Compute R^2 for the data of Example 7.1 and interpret the result.

Solution. From Example 7.2, SSE = 7.39 and SSyy = 973.5. Therefore:

$$R^2 = 1 - \frac{7.39}{973.5} = 1 - 0.0076 = 0.9924.$$

About 99.24% of the variation in examination marks is explained by the variation in hours studied. This is an exceptionally high R^2, meaning that hours of study is an excellent predictor of marks in this sample. Note: such a high R^2 is unusual in real social-science data and arises here partly because the example was constructed for illustrative simplicity.

7.8 CORRELATION ANALYSIS: THE PEARSON CORRELATION COEFFICIENT

While regression analysis models the direction and rate of the relationship between x and y, correlation analysis focuses on the strength and direction of the linear association between the two variables, without specifying which variable is the "cause" and which is the "effect". The standard measure of linear association is the Pearson product-moment correlation coefficient, universally denoted by r (in the sample) and ρ (rho, in the population).

7.8.1 The Sample Correlation Coefficient

For a sample of n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the sample correlation coefficient is defined by:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} \qquad (7.10)$$

An equivalent computing formula, in terms of the original summations, is:

$$r = \frac{n\,\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\left(n\,\Sigma x_i^2 - (\Sigma x_i)^2\right) \cdot \left(n\,\Sigma y_i^2 - (\Sigma y_i)^2\right)}} \qquad (7.11)$$

7.8.2 Properties of r

The correlation coefficient r has the following important properties:

1. r always lies between −1 and +1.
2. r = +1 if and only if all the sample points lie exactly on a straight line with a positive slope (a perfect positive linear relationship).
3. r = −1 if and only if all the sample points lie exactly on a straight line with a negative slope (a perfect negative linear relationship).
4. r = 0 indicates no linear association between x and y. (It does not mean there is no relationship at all — there could be a strong non-linear relationship.)

5. r has the same sign as the regression slope b_1: both are positive for a positive relationship and negative for a negative one.

6. In simple linear regression, $R^2 = r^2$. This identity connects the regression and correlation analyzes: the coefficient of determination is just the square of the correlation coefficient.

Example 7.4 (Pearson Correlation Coefficient)

Compute the sample correlation coefficient r for the data of Example 7.1 and verify that $R^2 = r^2$.

Solution. From the earlier calculations: $SSxy = 330.25$, $SSxx = \Sigma x^2 - (\Sigma x)^2/n = 1709 - (113)^2/8 = 1709 - 1596.125 = 112.875$, and $SSyy = 973.5$. Applying equation (7.10):

$$r = \frac{330.25}{\sqrt{112.875 \times 973.5}} = \frac{330.25}{\sqrt{109883.625}} = \frac{330.25}{331.49} = 0.9963.$$

Verification: $r^2 = (0.9963)^2 = 0.9926$, which agrees (within rounding error) with the $R^2 = 0.9924$ we computed in Example 7.3. The very high correlation of 0.9963 confirms the strong positive linear relationship between hours studied and examination marks.

Example 7.4(b) (A Complete Regression with a Negative Relationship)

A sociologist in Islamabad investigates the relationship between the unemployment rate (x, in per cent) and the average monthly household savings (y, in thousand rupees) for n = 7 administrative districts of a province. The data are as follows:

District	A	B	C	D	E	F	G
Unemployment (x)	4	6	8	5	10	7	12
Savings (y)	35	28	20	30	12	24	8

Solution. We construct the calculation table:

x	Y	xy	x^2	y^2
4	35	140	16	1225
6	28	168	36	784
8	20	160	64	400
5	30	150	25	900
10	12	120	100	144
7	24	168	49	576
12	8	96	144	64
$\Sigma = 52$	$\Sigma = 157$	$\Sigma = 1002$	$\Sigma = 434$	$\Sigma = 4093$

The sample means are $\bar{x} = 52/7 = 7.429$ and $\bar{y} = 157/7 = 22.429$.

$$b_1 = \frac{7(1002) - (52)(157)}{7(434) - (52)^2} = \frac{7014 - 8164}{3038 - 2704} = \frac{-1150}{334} = -3.443.$$

$$b_0 = 22.429 - (-3.443)(7.429) = 22.429 + 25.579 = 48.01.$$

The estimated regression equation is $\hat{y} = 48.01 - 3.443x$. The slope of −3.443 means that for each additional percentage point of unemployment, the predicted average monthly household savings decrease by about Rs. 3,443. This is a negative relationship, as expected: higher unemployment depresses savings.

Computing the correlation:

$$SS_{xx} = 434 - (52)^2/7 = 434 - 386.29 = 47.71$$

$$SS_{yy} = 4093 - (157)^2/7 = 4093 - 3521.29 = 571.71$$

$$SS_{xy} = 1002 - (52)(157)/7 = 1002 - 1166.29 = -164.29$$

$$r = \frac{-164.29}{\sqrt{47.71 \times 571.71}} = \frac{-164.29}{\sqrt{27275.49}} = \frac{-164.29}{165.15} = -0.995.$$

The very high negative correlation confirms the strong inverse linear relationship. $R^2 = (-0.995)^2 = 0.990$, meaning that about 99% of the variation in savings across districts is explained by the unemployment rate.

Testing the significance of the slope: SSE = 571.71 − (−3.443)(−164.29) = 571.71 − 565.59 = 6.12; $s_e = \sqrt{(6.12/5)} = \sqrt{1.224} = 1.107$.

$$S_{b_1} = \frac{1.107}{\sqrt{47.71}} = \frac{1.107}{6.907} = 0.1603.$$

$$t = \frac{-3.443}{0.1603} = -21.48.$$

With df = 5 and $t_{0.025,\,5} = 2.571$, $|t| = 21.48 \gg 2.571$, so the slope is overwhelmingly significant. The negative relationship between unemployment and savings is not a product of chance.

7.8.3 Interpreting the Magnitude of r

Students often ask what constitutes a "strong" correlation. There is no universal answer, because the strength depends on the field of study and the context of the research, but the following rough guidelines — adapted from Cohen (1988) — are widely used in the social sciences:

|r|	**Strength of Linear Association**
0.00 – 0.10	Negligible or trivial
0.10 – 0.30	Weak (small)
0.30 – 0.50	Moderate (medium)
0.50 – 0.70	Strong (large)
0.70 – 1.00	Very strong

These guidelines should be applied with caution. In physical sciences, correlations below 0.90 would be considered weak, while in social-science survey research a correlation of 0.40 or 0.50 can be very useful. The student should always consider the substantive context and the coefficient of determination R^2 (which tells what fraction of the variation is "explained") before making a judgement about practical importance.

7.9 INFERENCES ABOUT POPULATION REGRESSION PARAMETERS

Up to now we have been concerned only with the sample regression line $\hat{y} = b_0 + b_1 x$. But b_0 and b_1 are sample statistics, and like all sample statistics they are subject to sampling variability: a different sample would give different values. We now ask whether the relationship observed in the sample is "real" — that is, whether it also holds in the population — or whether it could plausibly have arisen by chance alone. The key inferential question in simple regression is almost always about the slope: does the population slope β_1 differ from zero?

If $\beta_1 = 0$, then the population regression line is flat: the value of x has no effect on the expected value of y, and the scatter diagram would show no linear trend. If $\beta_1 \neq 0$, then x and y are linearly related in the population, and the sign and magnitude of β_1 tell us the direction and rate of that relationship.

7.9.1 The ANOVA Approach to Testing the Overall Regression

Just as in the ANOVA of Unit 6, the total variability of y can be decomposed into two additive parts — a "regression" part (explained by x) and a "residual" part (unexplained):

$$SS_{yy} = SSR + SSE \qquad (7.12a)$$

where $SSR = b_1 SSxy$ is the regression sum of squares and $SSE = SSyy - SSR$ is the residual sum of squares. Dividing each by its degrees of freedom gives the mean squares:

$$MSR = \frac{SSR}{1} = SSR \;,\; MSE = \frac{SSE}{n-2} \qquad (7.12b)$$

The F-ratio $F = MSR/MSE$ follows the F-distribution with $(1, n-2)$ degrees of freedom under $H_0: \beta_1 = 0$. This F-test gives exactly the same result as the t-test for the slope (in fact, $F = t^2$ with 1 numerator df), but it is particularly useful in multiple regression where the overall F-test has more than one numerator degree of freedom. The regression ANOVA table looks very similar to the one-way ANOVA table of Unit 6:

Source	SS	df	MS	F
Regression	SSR	1	MSR	MSR/MSE
Residual	SSE	$n-2$	MSE	—
Total	SSyy	$n-1$	—	—

7.9.2 The Standard Error of the Slope

Under the LINE assumptions, the sampling distribution of the least-squares estimator b_1 is normal with mean β_1 and standard error:

$$SE_{b_1} = \frac{s_e}{\sqrt{SS_{xx}}} \qquad (7.13)$$

where s_e is the standard error of estimate (equation 7.7) and $SSxx = \Sigma(x_i - \bar{x})^2$ is the sum of squared deviations of x from its mean.

7.9.3 Test of Hypothesis for the Slope

The most common test is H_0: $\beta_1 = 0$ (no linear relationship) against H_1: $\beta_1 \neq 0$ (or the corresponding one-sided alternatives). The test statistic is:

$$t = \frac{b_1 - 0}{SE_{b_1}} = \frac{b_1}{s_e / \sqrt{SS_{xx}}} \qquad (7.14)$$

which follows Student's t-distribution with df = n − 2 under H_0. The decision rule is the same as in Unit 4: reject H_0 if $|t| > t_{\alpha/2,n-2}$ for a two-tailed test. Note that t^2 from equation (7.14) is exactly equal to the F from the regression ANOVA; the two approaches always give the same conclusion.

7.9.4 Confidence Interval for the Slope

The $(1 - \alpha) \times 100\%$ confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2,n-2} \cdot SE_{b_1} \qquad (7.15)$$

If the interval contains zero, the slope is not significantly different from zero at that confidence level, and we conclude that the evidence for a linear relationship is weak. If it does not contain zero, the relationship is significant, and the endpoints of the interval give the plausible range for the rate of change.

7.9.5 Test and Confidence Interval for the Correlation Coefficient

To test H_0: $\rho = 0$ against H_1: $\rho \neq 0$ (or the corresponding one-sided alternatives), the test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (7.23)$$

which also follows the t-distribution with df = n − 2. This test and the test for $\beta_1 = 0$ always give exactly the same result — the same t-value — because, in simple regression, the regression slope and the correlation coefficient are measuring the same thing from two different perspectives. This deep connection between regression and correlation is one of the most beautiful facts in introductory statistics.

Example 7.5 (Test of Significance for the Slope and Correlation)

For the data of Example 7.1, test at the 5% level of significance whether the population slope β_1 is significantly different from zero. Also test whether the population correlation ρ is significantly different from zero.

Solution. From the earlier calculations: $b_1 = 2.926$, $s_e = 1.110$, SSxx = 112.875, r = 0.9963, and n = 8.

Test of β_1:

$$SE_{b_1} = \frac{1.110}{\sqrt{112.875}} = \frac{1.110}{10.624} = 0.10448.$$

$$t = \frac{2.926}{0.10448} = 28.00.$$

With df = 8 − 2 = 6 and $\alpha = 0.05$, the critical value is $t_{0.025,\,6} = 2.447$. Since $28.00 \gg 2.447$, we reject H_0 overwhelmingly. The population slope is highly significantly different from zero.

Test of ρ:

$$t = \frac{0.9963 \times \sqrt{6}}{\sqrt{1 - 0.9926}} = \frac{0.9963 \times 2.449}{\sqrt{0.0074}} = \frac{2.440}{0.0860} = 28.37.$$

This is essentially the same value (the tiny difference is due to rounding). Both tests give exactly the same conclusion: the linear relationship between hours studied and marks is overwhelmingly significant.

7.10 PREDICTION USING THE REGRESSION LINE

One of the main practical uses of regression analysis is prediction: given a new value of x, what value of y should we expect? The point prediction is simply $\hat{y} = b_0 + b_1 x$. But a single number, without any indication of its reliability, is of limited use. We therefore construct a prediction interval — an interval estimate for a single future observation of y at a given value x_0. The $(1 - \alpha) \times 100\%$ prediction interval is:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \qquad (7.23)$$

There is also a narrower confidence interval for the mean of y at x_0 (rather than for a single future observation), which omits the "1 +" inside the square root. The prediction interval is wider than the mean confidence interval because predicting a single observation requires allowance for the individual randomness ε as well as for the uncertainty in the estimated line.

Example 7.6 (Point and Interval Prediction)

Using the regression equation from Example 7.1, (a) predict the marks of a student who studies $x_0 = 17$ hours per week, and (b) construct a 95% prediction interval for this student's marks.

Solution.

(a) $\hat{y}_0 = 22.42 + 2.926(17) = 22.42 + 49.74 = 72.16$ marks.

(b) We need $s_e = 1.110$, $\bar{x} = 14.125$, SSxx = 112.875, n = 8, and $t_{0.025,\,6} = 2.447$.

$$PI = 72.16 \pm 2.447 \times 1.110 \times \sqrt{1 + \frac{1}{8} + \frac{(17 - 14.125)^2}{112.875}}$$

$$= 72.16 \pm 2.447 \times 1.110 \times \sqrt{1 + 0.125 + 0.0732}$$

$$= 72.16 \pm 2.447 \times 1.110 \times \sqrt{1.1982}$$

$$= 72.16 \pm 2.447 \times 1.110 \times 1.0946 = 72.16 \pm 2.97.$$

The 95% prediction interval for the marks of a student who studies 17 hours per week is approximately (69.2 ; 75.1). We are 95% confident that this particular student's marks will fall between about 69 and 75.

7.11 LIMITATIONS, ERRORS AND CAVEATS OF REGRESSION AND CORRELATION

Regression and correlation are powerful tools, but they are also easily misused. The student who wishes to become a competent applied researcher must be at least as familiar with the warnings below as with the formulas themselves. In the words of the eminent statistician George E. P. Box, "all models are wrong, but some are useful". The simple linear regression model is useful precisely because it is simple, but that simplicity comes with a set of limitations that must always be kept in mind.

7.11.1 Correlation Does Not Imply Causation

This is perhaps the most important single warning in all of statistics. A high correlation between two variables means only that they tend to move together; it does not mean that one causes the other. The classic example is the positive correlation between the number of fire engines dispatched to a fire and the amount of damage caused by the fire: fires to which more engines are sent tend to be more destructive. But sending more fire engines does not cause more damage; both variables are caused by a third, unobserved variable — the severity of the fire. This phenomenon is called spurious correlation, and it is extremely common in social-science data where many variables are interrelated through complex webs of causation. Only a carefully controlled randomized experiment can establish causation; an observational correlation, however strong, cannot do so by itself.

A closely related problem is confounding. Two variables may be correlated because they are both related to a third variable (the confound) rather than because one directly affects the other. For example, ice-cream sales and drowning deaths are positively correlated, not because eating ice cream causes drowning, but because both are influenced by the weather: in hot weather people eat more ice cream and also swim more. Confounding is pervasive in observational studies and is one of the main reasons that social scientists must be extremely cautious when interpreting regression coefficients as "effects" of one variable on another.

7.11.2 Do Not Extrapolate Beyond the Range of the Data

The regression equation is an estimate of the relationship between x and y within the range of x-values actually observed in the sample. Using the equation to predict y for an x-value that lies far outside this range is called extrapolation, and it is extremely risky because there is no guarantee that the linear relationship continues to hold outside the observed range. A regression line fitted to the relationship between income and expenditure for households earning Rs. 20,000 to Rs. 80,000 per month should not be used to predict the expenditure of a household earning Rs. 500,000 — the relationship may be quite different at that level of income. The danger of extrapolation increases rapidly as the prediction moves further from the range of the data.

7.11.3 Beware of Influential Observations and Outliers

A single unusual observation — one that is very far from the centre of the data in either the x-direction or the y-direction — can have a dramatic effect on the slope and the intercept of the regression line, and on the correlation coefficient. Such observations are called influential points or leverage points, and they should always be identified and examined carefully. A scatter diagram makes influential points easy to spot; they are the points that stand far away from the main cluster. The researcher should never simply delete an outlier without a good reason (such as a confirmed recording error), but should always investigate whether the regression results change substantially when the outlier is included and when it is excluded. If a single observation changes the conclusion dramatically, the conclusion rests on very fragile ground.

7.11.4 r ≈ 0 Does Not Mean No Relationship

The Pearson correlation coefficient measures only linear association. Two variables may have a strong, highly systematic non-linear relationship — a U-shape, a sine wave, a quadratic curve — and still have a correlation coefficient close to zero. The scatter diagram is the safeguard against this misinterpretation: always look at the scatter diagram before computing r. If the scatter diagram reveals a non-linear pattern, a linear regression and a Pearson correlation are the wrong tools for the job, and a non-linear model or a rank-based measure of association should be considered instead.

7.11.5 Ecological Fallacy

The ecological fallacy is the error of drawing conclusions about individuals from data that have been aggregated at the group level. If a regression of average literacy rate on average per capita income across 50 districts gives $R^2 = 0.85$, it does not follow that the same strong relationship holds at the individual level — the correlation between an individual person's income and his or her literacy status may be much weaker. Aggregated data tend to smooth out individual variation and thereby inflate R^2, sometimes dramatically. The student should always be clear about the unit of analysis (individuals, households, districts, countries) and should not generalize from one level to another without strong justification.

7.11.6 Regression Towards the Mean

Sir Francis Galton, the inventor of regression analysis, noticed a curious phenomenon when studying the heights of fathers and sons: the sons of very tall fathers tended to be tall, but not as tall as their fathers; the sons of very short fathers tended to be short, but not as short as their fathers. In both cases, the sons' heights "regressed towards" the mean of the population. This phenomenon — which Galton called "regression to the mean" and which gave the technique its name — is a purely statistical effect, not a biological or causal one. It arises whenever two measurements are imperfectly correlated: extreme values on the first measurement tend to be associated with less extreme values on the second, purely because of the element of randomness. Regression towards the mean is a common source of confusion in before-and-after studies and in repeated measurements.

7.11.7 The Spearman Rank Correlation Coefficient

When the data are ordinal (ranked) rather than interval, or when the relationship is monotonic but not necessarily linear, the Pearson correlation coefficient r may give a misleading picture of the strength of the association. In such situations the Spearman rank correlation coefficient, denoted r_s, is a useful alternative. The Spearman coefficient is simply the Pearson correlation applied to the ranks of the observations rather than to the observations themselves:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (7.17a)$$

where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ is the difference between the ranks of the i-th observation on the two variables, and n is the number of paired observations. The Spearman coefficient lies between −1 and +1, with the same interpretation as the Pearson coefficient but without the assumption of linearity or interval-level measurement. When there are no tied ranks, equation (7.17a) gives exactly the same result as computing the Pearson correlation on the ranked data.

Example 7.5(b) (Spearman Rank Correlation)

Six students are ranked on two variables: hours of study per week and examination marks. The ranks are as follows:

Student	1	2	3	4	5	6
Rank in hours (R_1)	3	1	5	2	6	4
Rank in marks (R_2)	2	1	6	3	5	4

Solution. Compute the rank differences d and their squares:

Student	1	2	3	4	5	6	Total
$d = R_1 - R_2$	1	0	−1	−1	1	0	0
d^2	1	0	1	1	1	0	4

Applying equation (7.17a):

$$r_s = 1 - \frac{6(4)}{6(36 - 1)} = 1 - \frac{24}{210} = 1 - 0.114 = 0.886.$$

The Spearman rank correlation of 0.886 indicates a very strong positive monotonic association between hours studied and examination marks in this sample. The advantage of the Spearman coefficient over the Pearson coefficient in this context is that it makes no assumption about the form of the relationship — it only requires that higher hours tend to go with higher marks.

7.12 MULTIPLE REGRESSION AND CORRELATION ANALYSIS

In most real social-science research, the value of a dependent variable depends not on one but on several independent variables simultaneously. A student's examination marks depend not only on hours of study but also on prior academic ability, motivation, quality of teaching, access to resources and many other factors. A household's monthly food expenditure depends not only on income but also on household size, location, educational level and dietary preferences. To model such situations we need multiple regression, which extends simple regression by including two or more independent variables in the same equation.

The general multiple regression model with p independent variables is:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \qquad (7.23)$$

where x_{1i}, x_{2i}, …, x_{pi} are the values of the p independent variables for observation i, β_0, β_1, …, β_p are the unknown regression parameters, and ε_i is the random error with the same assumptions as in simple regression (independence, zero mean, constant variance σ^2, and normality).

The parameter β_j (j = 1, 2, …, p) is called the partial regression coefficient for variable x_j; it represents the expected change in Y for a one-unit increase in x_j, holding all the other independent variables constant. This "holding everything else constant" interpretation is the key difference between simple and multiple regression, and it is one of the reasons why multiple regression is so much more useful than running a series

of separate simple regressions — it allows us to disentangle the individual effects of several correlated variables.

7.13 FINDING THE MULTIPLE REGRESSION EQUATION

The sample multiple regression equation is written:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (7.23)$$

where b_0, b_1, …, b_p are the least-squares estimates of the corresponding population parameters. These estimates are obtained by minimizing the sum of squared residuals $\Sigma(y_i - \hat{y}_i)^2$, exactly as in simple regression. For $p \geq 2$ the resulting system of normal equations is most conveniently written in matrix form ($Y = X\beta + \varepsilon$, with the least-squares solution $b = (X'X)^{-1}X'Y$), and the actual computation is almost always done by computer software such as Excel, SPSS, R or Minitab. We shall not present the matrix derivation here; instead, we illustrate the procedure with a two-variable example in which the computation can be done by hand.

Example 7.7 (Multiple Regression with Two Independent Variables)

A social scientist wishes to model the monthly household expenditure on food (y, in thousand rupees) as a function of monthly income (x_1, in thousand rupees) and household size (x_2, number of persons). A random sample of n = 10 households in Rawalpindi gives:

HH	Y	x_1	x_2
1	15	40	3
2	18	50	4
3	22	55	5
4	12	35	2
5	25	70	6
6	20	60	4
7	28	80	5
8	16	45	3
9	30	90	7
10	24	65	5

Using statistical software (or, equivalently, solving the normal equations by hand), the estimated multiple regression equation is found to be:

$$\hat{y} = -1.82 + 0.22x_1 + 1.95x_2$$

Interpretation. For each additional thousand rupees of monthly income (holding household size constant), food expenditure is estimated to increase by about Rs. 220. For each additional person in the household (holding income constant), food expenditure is estimated to increase by about Rs. 1,950. Both partial effects are intuitively sensible: higher-income households spend more on food, and larger households also spend more.

The multiple coefficient of determination is $R^2 = 0.984$, meaning that about 98.4% of the variation in food expenditure is jointly explained by income and household size — a very high value for social-science data.

7.14 INFERENCES IN MULTIPLE REGRESSION

7.14.1 The Overall F-Test

The first inferential question in multiple regression is whether the set of independent variables, taken together, has any explanatory power at all. The hypotheses are:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad versus \quad H_1: at\ least\ one\ \beta_j \neq 0 \qquad (7.23)$$

The test statistic is:

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \qquad (7.23)$$

which follows an F-distribution with $\nu_1 = p$ numerator and $\nu_2 = n - p - 1$ denominator degrees of freedom under H_0. If F exceeds the critical value $F_{\alpha,\, p,\, n-p-1}$, we conclude that the regression as a whole is statistically significant. In Example 7.7, with $R^2 = 0.984$, the overall F-test would reject overwhelmingly at any conventional α.

7.14.2 Individual t-Tests for Each Coefficient

Even if the overall F-test is significant, it is possible that some individual independent variables do not contribute significantly to the regression once the others are already in the model. To test whether a particular β_j is significantly different from zero, the test statistic is:

$$t = \frac{b_j}{SE_{b_j}} \qquad (7.23)$$

with $df = n - p - 1$. This is the two-tailed t-test, exactly analogous to the test for the slope in simple regression (equation 7.13). The standard error SE_{b_j} depends on both the residual variance and the correlations among the independent variables; it is routinely printed by all standard statistical software.

7.14.3 The Adjusted R^2

When independent variables are added to a regression model, R^2 can only increase or stay the same — it never decreases. This is misleading, because adding a useless variable can increase R^2 slightly purely by chance. The adjusted R^2, denoted $\bar{R}^2$, penalises the model for each additional independent variable and is defined by:

$$R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \qquad (7.23)$$

$\bar{R}^2$ can actually decrease when a useless variable is added, so it is a more honest measure of the model's explanatory power than R^2 itself. The student should always report $\bar{R}^2$ in addition to R^2 when presenting the results of a multiple regression.

7.15 MODELLING TECHNIQUES: VARIABLE SELECTION AND DIAGNOSTICS

Building a good multiple regression model is as much an art as a science, and it involves a number of choices that cannot be made by formula alone. We discuss three of the most important aspects briefly; a full treatment would require a separate course.

7.15.1 Variable Selection

When many potential independent variables are available, the researcher must decide which ones to include in the model. Including too few can lead to omitted-variable bias (the regression attributes to the included variables an effect that really belongs to the omitted ones); including too many can lead to overfitting (the model captures noise rather than signal, and the estimates become imprecise because of multicollinearity). Several formal procedures for variable selection are available, including forward selection (start with no variables and add the most significant one at each step), backward elimination (start with all variables and remove the least significant one at each step), and stepwise regression (a combination of both). These procedures are implemented in all standard software packages and are extremely widely used, although they have well-known theoretical weaknesses and should be used with caution.

7.15.2 Multicollinearity

Multicollinearity is the phenomenon in which two or more independent variables are highly correlated with one another. When multicollinearity is severe, the individual regression coefficients become very unstable (their standard errors become very large), making it difficult to disentangle the separate effects of the correlated variables even though the overall regression may fit the data well (R^2 is high). A common diagnostic for multicollinearity is the variance inflation factor (VIF) for each independent variable; a VIF greater than about 5 or 10 is usually taken as a signal of concern. Remedies include dropping one of the correlated variables, combining them into a single index, or using a specialized technique such as ridge regression.

7.15.3 Residual Diagnostics

After fitting a regression model, the researcher should always examine the residuals (the differences $e_i = y_i - \hat{y}_i$) to check whether the LINE assumptions are satisfied. A plot of the residuals against the fitted values $\hat{y}$ should show a random scatter with no clear pattern; a funnel shape (residuals becoming larger for larger $\hat{y}$) suggests heteroscedasticity, and a curve suggests non-linearity. A normal probability plot (Q–Q plot) of the residuals should be approximately a straight line if the normality assumption holds. Outlying residuals (large in absolute value) may indicate influential observations that need special attention. All standard software packages produce these diagnostic plots automatically, and the student is strongly encouraged to inspect them routinely before trusting any regression result.

7.16 SUMMARY

This unit has developed the theory and practice of regression and correlation analysis, two of the most widely used statistical techniques in the social sciences. The main points may be summarized as follows.

- A scatter diagram is an indispensable first step in any analysis of the relationship between two quantitative variables. It reveals the direction, the form and the strength of the relationship at a glance, and it warns the researcher immediately if the linearity assumption is violated.
- The simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$ describes the relationship between a dependent variable Y and a single independent variable x. The parameters β_0 (intercept) and β_1 (slope) are

estimated from the sample by the method of least squares, which minimizes the sum of squared residuals. The resulting sample regression line $\hat{y} = b_0 + b_1 x$ always passes through the point $(\bar{x}, \bar{y})$.

- The slope b_1 is interpreted as the estimated change in y for each one-unit increase in x. It is the most important quantity in a simple regression analysis. The intercept b_0 is the predicted value of y when $x = 0$, and may or may not have a meaningful practical interpretation depending on the context.
- The standard error of estimate s_e measures the average size of the residuals and is computed with $(n - 2)$ degrees of freedom. The coefficient of determination R^2 measures the proportion of the total variation in y that is explained by the regression on x. In simple regression, $R^2 = r^2$, where r is the Pearson correlation coefficient.
- The Pearson correlation coefficient r measures the strength and direction of the linear association between x and y. It always lies between −1 and +1, with ±1 indicating a perfect linear relationship and 0 indicating no linear association. A correlation of zero does not imply no relationship — only no linear relationship.
- Inferences about the population slope β_1 and the population correlation ρ use the t-distribution with $df = n - 2$. The test of H_0: $\beta_1 = 0$ and the test of H_0: $\rho = 0$ always give exactly the same result. Confidence intervals and prediction intervals for individual observations are also based on the t-distribution and use the standard error of estimate.
- Correlation does not imply causation. Extrapolation beyond the range of the data is risky. Influential observations can dramatically distort regression results. A correlation of zero does not mean no relationship — only no linear relationship. These four warnings should be kept in mind in every regression analysis.
- Multiple regression extends simple regression to the case of two or more independent variables. The partial regression coefficient β_j represents the expected change in Y for a one-unit increase in x_j, holding all other variables constant. The overall F-test checks whether the regression as a whole is significant; individual t-tests check whether particular variables contribute significantly.
- The adjusted R^2 penalises the model for unnecessary variables and is a more honest measure of explanatory power than R^2 alone. Modelling techniques such as variable selection, multicollinearity diagnostics (VIF) and residual analysis are essential complements to the numerical output of any regression.

In Unit 8 we shall develop a set of non-parametric methods that can be used when the assumptions of the parametric techniques (normality, linearity, constant variance) are not satisfied. In Unit 9 we shall study time-series analysis and forecasting, which extend regression ideas to data collected over time. The student who has mastered the material of the present unit will find both of these later units a natural continuation.

7.17 SELF-ASSESSMENT EXERCISE

1. Distinguish between regression analysis and correlation analysis. What question does each one answer? Can one be used without the other?
2. State the four assumptions (LINE) of the simple linear regression model. For each assumption, briefly describe what happens to the regression results if it is violated.
3. What is the interpretation of the slope b_1 in a simple linear regression equation? What is the interpretation of the intercept b_0? Under what circumstances is the intercept not practically meaningful?

4. Explain the difference between the coefficient of determination R^2 and the adjusted R^2. Why is the adjusted R^2 preferred in multiple regression?

5. Why does "correlation does not imply causation"? Give an example of two variables that are highly correlated but for which neither is the cause of the other.

6. Derive the least-squares formulas for the slope b_1 and the intercept b_0 of the simple linear regression line $\hat{y} = b_0 + b_1x$. Begin by writing down the sum of squared residuals $S = \Sigma(y_i - b_0 - b_1x_i)^2$, take partial derivatives with respect to b_0 and b_1, set them equal to zero, and solve for b_1 and b_0. Show that the resulting line always passes through the point $(\bar{x}, \bar{y})$. Illustrate with a complete numerical example of your own choosing.

7. Discuss the coefficient of determination R^2, the standard error of estimate s_e, and the Pearson correlation coefficient r as alternative measures of the quality of a regression fit. For each measure, give its definition, its range of possible values, and its interpretation. Explain the relationship $R^2 = r^2$ in simple regression, and discuss why this relationship does not hold in multiple regression.

8. A simple regression analysis produces $b_1 = 3.4$ with $SE_{b_1} = 1.8$ and $n = 12$ observations. (a) Without computing the t-statistic, decide whether the slope is likely to be significant at $\alpha = 0.05$ (two-tailed). Explain your reasoning. (b) Now compute the t-statistic and the exact decision. (c) Suppose the sample size were doubled to $n = 24$ while b_1 and s_e remained the same. Would the t-statistic change? In which direction and why? [Answers: (a) $t \approx 1.89$, critical $t_{0.025,\ 10} = 2.228 \rightarrow$ not significant; (b) confirms; (c) SE would decrease, so t would increase — larger samples give more power.]

9. A researcher reports $R^2 = 0.85$ and adjusted $R^2 = 0.60$ for a regression with $p = 12$ independent variables and $n = 20$ observations. (a) Explain why there is such a large gap between R^2 and adjusted R^2. (b) What advice would you give the researcher about the model? (c) If the researcher were to drop 8 of the 12 variables that contribute very little, what would you expect to happen to R^2 and to adjusted R^2?

10. The correlation between the number of smartphones sold and the number of hospital admissions in a country over a 20-year period is $r = 0.95$. A journalist writes: "Smartphones cause hospital admissions." Discuss, in the language of this unit, what is wrong with this conclusion and propose a more careful interpretation.

11. A simple regression gives $\hat{y} = 10 + 2x$, fitted to data with x ranging from 5 to 30. A manager asks you to predict y for $x = 80$. Explain why this request is problematic and what you would advise.

12. Two simple regressions are fitted to the same sample of $n = 50$ observations. Model A uses x_1 as the independent variable and gives $R^2 = 0.45$. Model B uses x_2 and gives $R^2 = 0.38$. Can you conclude that x_1 is a "better" predictor than x_2 in a multiple regression that includes both? Why or why not?

13. The following data represent the age (x, in years) and the systolic blood pressure (y, in mmHg) of $n = 10$ patients at a clinic in Islamabad:

Age (x)	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**
BP (y)	118	122	126	132	138	140	148	152	160	165

(a) Compute b_1, b_0 and write the regression equation. (b) Compute R^2 and r. (c) Interpret the slope. (d) Predict the BP of a 48-year-old patient. [Answers: $\Sigma x = 475$, $\Sigma y = 1401$, $\Sigma xy = 68345$, $\Sigma x^2 = 24125$, $\Sigma y^2 = 198849$; $b_1 \approx 1.013$; $b_0 \approx 92.01$; $\hat{y} = 92.01 + 1.013x$; $R^2 \approx 0.989$; $r \approx 0.995$; predicted BP at 48 $\approx$ 140.6 mmHg]

14. Using the data from the previous problem: (a) Compute the standard error of estimate s_e. (b) Test at $\alpha = 0.01$ whether the slope is significantly different from zero. (c) Construct a 95% confidence

interval for the population slope β_1. [Answers: $s_e \approx 1.71$; $SE_{b_1} \approx 0.0425$; $t \approx 23.8$; df = 8; reject very strongly; 95% CI for $\beta_1 \approx (0.915 ; 1.111)$]

15. The following data show the weekly advertising expenditure (x, in thousand rupees) and the weekly sales revenue (y, in thousand rupees) for a small firm in Lahore over 8 weeks:

Advertising (x)	**5**	**8**	**10**	**12**	**6**	**9**	**11**	**7**
Sales (y)	50	65	80	90	55	70	85	60

(a) Fit a simple linear regression of y on x. (b) Compute R^2 and s_e. (c) Test at $\alpha = 0.05$ whether advertising expenditure significantly affects sales. (d) Predict the sales revenue if the firm spends Rs. 15,000 on advertising in a particular week, and comment on the reliability of this prediction. [Answers: $b_1 \approx 5.49$; $b_0 \approx 20.43$; $\hat{y} = 20.43 + 5.49x$; $R^2 \approx 0.987$; $s_e \approx 2.02$; $t \approx 17.4$; reject; $\hat{y}(15) \approx 102.8$ — but 15 is well outside the observed range of x (5 to 12), so this is extrapolation and should be treated with caution.]

16. For the advertising data of the previous problem, (a) compute the Pearson correlation coefficient r and verify that $R^2 = r^2$. (b) Test at $\alpha = 0.05$ whether the population correlation ρ is significantly different from zero using equation (7.15). [Answers: $r \approx 0.994$; $r^2 = 0.988 \approx R^2$; $t \approx 17.7$; reject at any conventional level.]

17. A researcher at AIOU collected data on the number of assignments submitted (x) and the final examination score (y, out of 100) for n = 12 distance-learning students. The summary statistics are: $\Sigma x = 72$, $\Sigma y = 756$, $\Sigma x^2 = 480$, $\Sigma y^2 = 48960$, $\Sigma xy = 4740$. (a) Compute b_1, b_0 and write the regression equation. (b) Compute R^2 and interpret it. [Answers: $\bar{x} = 6$, $\bar{y} = 63$; SSxx = 48, SSyy = 1332, SSxy = 228; $b_1 = 4.75$; $b_0 = 34.5$; $\hat{y} = 34.5 + 4.75x$; $R^2 \approx 0.814$; about 81.4% of the variation in examination scores is explained by assignment submission.]

18. In Example 7.7 (multiple regression), the researcher is told that Adjusted $R^2 = 0.979$. (a) Verify this value using equation (7.22), given $R^2 = 0.984$, n = 10 and p = 2. (b) Compute the overall F-statistic using equation (7.20) given SSR = 302.76 and SSE = 4.92. (c) Is the overall regression significant at $\alpha = 0.01$? [Answers: (a) $\bar{R}^2 = 1 - (1 - 0.984)(9/7) = 1 - 0.0206 = 0.979$ ✓; (b) $F = (302.76/2)/(4.92/7) = 151.38/0.7029 = 215.4$; (c) $F_{0.01,\ 2,\ 7} \approx 9.55$; $F = 215.4 \gg 9.55$; reject overwhelmingly.]

19. A sociologist finds a Pearson correlation of $r = -0.42$ between the average class size and the percentage of students passing a standardized test, based on data from n = 30 schools. (a) Test at $\alpha = 0.05$ whether this correlation is significantly different from zero. (b) Compute R^2 and comment on the practical importance of the relationship. [Answers: (a) $t = -0.42\sqrt{28}/\sqrt{(1-0.1764)} = -0.42(5.292)/0.9075 = -2.449$; df = 28; $|t| > 2.048 = t_{0.025,\ 28}$; reject; (b) $R^2 = 0.1764 \approx 17.6\%$; statistically significant but only 17.6% of variation is explained — the practical importance is modest.]

20. A quality engineer at a factory in Sialkot records the speed (x, in rpm) and the number of defective items per hour (y) on a production line over 10 shifts:

Speed (x)	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**	**280**
Defects (y)	3	5	4	8	7	10	12	9	14	16

(a) Fit a simple linear regression. (b) Compute R^2 and test the significance of the slope at $\alpha = 0.05$. (c) Comment on the practical implication for the factory manager. [Answers: $\Sigma x = 1900$, $\Sigma y = 88$; $b_1 \approx 0.0667$; $b_0 \approx -3.87$; $\hat{y} = -3.87 + 0.0667x$; $R^2 \approx 0.909$; $t \approx 8.46$; reject; for each additional 10 rpm, about 0.67 more defective items per hour are expected — the factory manager should consider running the line at a lower speed if quality is a priority.]

21. A market researcher collects data on the monthly expenditure on food (y) as a function of income (x_1) and household size (x_2) for 15 households. A multiple regression produces: $\hat{y} = 5.2 + 0.18x_1 + 2.1x_2$, $R^2 = 0.88$, $\bar{R}^2 = 0.86$, $s_e = 3.1$, and the individual t-values are $t_{b_1} = 4.5$ and $t_{b_2} = 3.2$ (both with df = 12). (a) Is the overall regression significant at $\alpha = 0.01$? (b) Is x_1 individually significant at $\alpha = 0.05$? (c) Is x_2 individually significant at $\alpha = 0.05$? (d) Interpret the two partial regression coefficients. [Answers: (a) $F \approx (0.88/2)/((1-0.88)/12) \times 15$−factor → use $F = [R^2/p]/[(1-R^2)/(n-p-1)] = (0.44)/(0.01) = 44.0$; $F_{0.01,2,12} = 6.927$; reject; (b) $t = 4.5 > t_{0.025,\,12} = 2.179$; significant; (c) $t = 3.2 > 2.179$; significant; (d) $b_1 = 0.18$: for each Rs. 1,000 increase in income, food expenditure rises by Rs. 180 holding household size constant; $b_2 = 2.1$: each additional person adds Rs. 2,100 to food expenditure holding income constant.]

22. A study of n = 40 firms shows $R^2 = 0.35$ for the regression of profit on advertising expenditure. A second study of n = 200 firms on the same variables shows $R^2 = 0.08$. (a) Which R^2 is "statistically more significant"? Why? (b) Which R^2 shows a stronger relationship in practical terms? (c) What does this example illustrate about the distinction between statistical significance and practical importance? [Answer: the second study has n = 200 and even a small R^2 is likely significant with such a large n; the first study has the stronger practical relationship (35% vs 8% of variation explained); this illustrates that large n makes even trivially small effects statistically significant.]

23. A researcher has the following residual plot from a regression analysis: the residuals fan out as the predicted values increase — that is, the scatter of residuals is narrow at the left of the plot and wide at the right. (a) Which of the LINE assumptions is being violated? (b) Name two possible remedies. [Answer: (a) Equal variance (homoscedasticity) is violated — the residuals show heteroscedasticity. (b) Possible remedies: (i) apply a variance-stabilising transformation such as log(y); (ii) use weighted least squares.]

UNIT 8: NONPARAMETRIC METHODS

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Explain why nonparametric methods are sometimes preferred over parametric tests, and identify situations (violation of normality, small samples, ordinal data, presence of outliers) in which they are most appropriate.
2. Apply the sign test to test a hypothesis about a population median (one-sample) and to compare two paired samples, and understand the binomial logic underlying the test.
3. Use the Wilcoxon signed-rank test for paired data and the Mann-Whitney U test for two independent samples, derive their mean and variance under the null hypothesis, and understand how these rank-based tests relate to the paired t-test and the two-sample t-test.
4. Carry out the Kruskal-Wallis H test to compare three or more independent populations, interpret its results as an alternative to one-way ANOVA, and perform post-hoc pairwise comparisons when the overall test is significant.
5. Perform the one-sample runs test to assess whether a sequence of binary outcomes is random, and compute the Spearman rank correlation coefficient as a nonparametric measure of association between two ordinal variables.
6. Recognize the Kolmogorov-Smirnov test as a method for comparing an empirical distribution with a theoretical distribution or for comparing two empirical distributions, and understand its advantages over the chi-square goodness-of-fit test.
7. Compare parametric and nonparametric tests in terms of asymptotic relative efficiency, power under non-normality, and robustness to outliers.

8.1 INTRODUCTION

Throughout the previous seven units we have relied heavily on the assumption that the populations we study are approximately normal, or that the sample sizes are large enough for the Central Limit Theorem to guarantee approximate normality of the sample means. The t-tests, the F-tests, and the confidence intervals of Units 4, 5, 6 and 7 are called parametric procedures because they make specific assumptions about the form of the underlying population distribution (most often normality) and about parameters such as the mean and variance. These assumptions, when satisfied, allow us to derive exact sampling distributions (t, F, chi-square) and to construct tests and confidence intervals with optimal properties.

But what happens when the normality assumption is clearly violated? What if the data are highly skewed, contain many outliers, or are measured only on an ordinal scale (such as "strongly disagree" to "strongly agree")? What if the sample sizes are so small that we cannot rely on the Central Limit Theorem? In such situations, parametric tests may give misleading results — the actual Type I error rate may be very different from the chosen α, and the power of the test may be severely reduced. For example, a two-sample t-test applied to data from heavy-tailed distributions can have a Type I error rate as high as 0.10 when α is set to 0.05, because the test statistic does not follow the t-distribution when the underlying population is not normal.

Nonparametric methods (also called distribution-free methods) come to the rescue. These procedures do not assume that the data come from any particular parametric family (such as the normal distribution). Instead, they are based on relatively weak assumptions — often only that the observations are independent and that the underlying distribution is continuous. Many nonparametric tests work by converting the original observations into ranks (their positions in an ordered list) and then analyzing the ranks rather than the raw data. Because ranks discard the exact numerical values but preserve the order, rank-based tests are resistant to outliers and do not require normality.

This unit introduces the most important nonparametric methods used in social-science research. We begin with the simplest — the sign test — which uses only the signs of differences. We then move to more powerful rank-based tests: the Wilcoxon signed-rank test (the nonparametric alternative to the paired t-test), the Mann-Whitney U test (the alternative to the two-sample t-test), and the Kruskal-Wallis H test (the alternative to one-way ANOVA). We also study the runs test for randomness, the Spearman rank correlation coefficient (the nonparametric alternative to Pearson's r), and the Kolmogorov-Smirnov test for comparing distributions. Each test is presented with its assumptions, its test statistic, its sampling distribution under the null hypothesis, and at least one worked example in a Pakistani context.

The student is strongly advised to keep Units 4, 5, 6 and 7 close at hand, because the logic of hypothesis testing (the six-step procedure) is exactly the same here. The only difference is the test statistic and the way the p-value is computed. As we proceed, we will repeatedly refer back to the parametric analogues, comparing their assumptions and relative efficiencies.

8.2 WHEN ARE NONPARAMETRIC METHODS NEEDED?

Before diving into the specific tests, it is useful to understand the situations in which a nonparametric test is preferable to its parametric counterpart. The following conditions should prompt the researcher to consider a nonparametric approach.

8.2.1 Severe violation of normality

The t-test and ANOVA are reasonably robust to moderate departures from normality, especially when sample sizes are large (say, $n > 30$). However, when the population distribution is strongly skewed, has heavy tails

(e.g., Cauchy or t-distribution with few degrees of freedom), or is clearly multimodal, and the sample size is small (say, $n < 30$), parametric tests can be unreliable. Nonparametric tests do not assume any particular shape of the distribution. They are based on the order of the observations, not their exact values, so they are unaffected by skewness or kurtosis.

8.2.2 Presence of outliers

Outliers have a disproportionate influence on the sample mean and the sample variance, and therefore on t-tests and F-tests. A single extreme observation can pull the mean away from the centre of the bulk of the data, inflate the variance, and lead to a non-significant result when a real effect exists (or a significant result when one does not). Because rank-based tests replace the actual values by their ranks, outliers are "flattened" — the largest observation receives the highest rank regardless of how extreme it is, and the smallest receives rank 1 regardless of how extreme it is. This makes rank tests very robust to outliers.

8.2.3 Ordinal (ranked) data

Many social-science measurements are inherently ordinal. For example, survey responses on a Likert scale ("poor", "fair", "good", "excellent") are ordered categories, but the intervals between categories are not necessarily equal. The mean of such data is not meaningful (it assumes equal spacing), but the median is. Nonparametric tests that work with ranks are perfectly appropriate for ordinal data, whereas parametric tests are not. Similarly, in educational research, class ranks, letter grades (A, B, C, D, F), and socioeconomic status categories (low, medium, high) are ordinal and should be analyzed with nonparametric methods.

8.2.4 Very small sample sizes

When n is very small (e.g., 5 or 6 per group), the Central Limit Theorem cannot be relied upon, and the normality assumption is impossible to verify. There is simply not enough data to test whether the population is normal. In such cases, nonparametric tests provide exact p-values (based on permutations of ranks) that do not require large-sample approximations. For example, with $n = 5$ in each group, the Mann-Whitney U test can enumerate all possible assignments of ranks and compute the exact p-value without any approximation.

8.2.5 When the parameter of interest is the median rather than the mean

For skewed distributions, the median is often a better measure of central tendency than the mean. Consider household income in Pakistan: a few very wealthy individuals can pull the mean far to the right, while the median remains representative of the typical household. Many nonparametric tests are naturally formulated as tests about the population median (sign test) or about the equality of distributions (which implies equality of medians if the distributions have the same shape).

8.2.6 When the population variance is not constant (heteroscedasticity)

The two-sample t-test (pooled version) assumes equal variances. Welch's t-test relaxes this assumption but still assumes normality. Nonparametric tests like the Mann-Whitney U test do not require equal variances; they test whether one distribution is stochastically larger than the other, regardless of the spread.

8.2.7 Relationship between nonparametric and parametric tests

The student should note that when the assumptions of the parametric test are satisfied, the parametric test is generally more powerful (more likely to detect a real effect). However, the loss of power from using a nonparametric test when normality holds is surprisingly small — often only about 5% for the Wilcoxon tests compared to the t-test. This is measured by the asymptotic relative efficiency (ARE). For the Wilcoxon

signed-rank test versus the t-test, the ARE is $3/\pi \approx 0.955$; for the Mann-Whitney versus the two-sample t-test, the ARE is also about 0.955. Therefore, in many practical situations, the nonparametric test is a safe and robust choice. In contrast, when normality is violated, the parametric test can have ARE approaching zero (i.e., it completely fails), while the nonparametric test maintains its power.

The following table summarizes the correspondence, which we will develop in detail throughout this unit:

Parametric test	Nonparametric alternative	When to prefer nonparametric
One-sample t-test (mean)	Sign test (median) or Wilcoxon signed-rank test	Skewed data, outliers, ordinal scale
Paired t-test	Sign test (paired) or Wilcoxon signed-rank test	Differences not normal, small n
Two-sample t-test (independent)	Mann-Whitney U test (Wilcoxon rank-sum test)	Unequal variances, non-normality, outliers
One-way ANOVA (F-test)	Kruskal-Wallis H test	Non-normal residuals, heterogeneous variances
Pearson correlation (r)	Spearman rank correlation (r_s)	Non-linear monotonic relationship, ordinal data
Goodness-of-fit (chi-square)	Kolmogorov-Smirnov (one-sample)	Continuous data, no arbitrary binning

8.3 THE SIGN TEST FOR A POPULATION MEDIAN

The sign test is the simplest of all nonparametric tests. Its origins go back to John Arbuthnot in 1710, who used a sign-like test to examine whether more boys than girls were born in London. The test is called the "sign" test because it replaces each observation by a sign (plus or minus) indicating whether the observation is above or below the hypothesized median. It is used to test a hypothesis about the median of a single population, and it requires no assumptions about the shape of the distribution — only that the variable is continuous (so that ties at the hypothesized median are rare) and that the observations are independent.

8.3.1 Derivation from the Binomial Distribution

Let η denote the population median. By definition, exactly half of the population values are less than η and half are greater than η, assuming a continuous distribution. Therefore, if we take a random sample of size n, the number of observations that fall below η follows a binomial distribution with parameters n and $p = 0.5$. Now suppose we want to test H_0: $\eta = \eta_0$ against H_1: $\eta \neq \eta_0$. Under H_0, the probability that any given observation is less than η_0 is exactly 0.5. Hence, the number X of observations less than η_0 (or, equivalently, greater than η_0) is Binomial(n, 0.5). If X is very small or very large, we have evidence against H_0.

Formally, define:

- n = sample size after discarding any observations exactly equal to η_0.
- X^+ = number of observations greater than η_0.
- X^- = number of observations less than $\eta_0 = n - X^+$.

Under H_0, $X^+ \sim$ Binomial(n, 0.5). The p-value for a two-tailed test is:

$$p = 2 \times min(P(X^+ \leq x^+), P(X^+ \geq x^+))$$

where x^+ is the observed number of values greater than η_0. For a right-tailed test H_1: $\eta > \eta_0$, $p = P(X^+ \geq x^+)$. For a left-tailed test H_1: $\eta < \eta_0$, $p = P(X^+ \leq x^+)$.

8.3.2 Handling Ties and Zero Differences

If any observation is exactly equal to η_0, it provides no information about whether the median is above or below η_0. The standard procedure is to discard such observations and reduce the sample size n accordingly. This is conservative because discarding ties reduces power, but it is necessary to preserve the binomial distribution with p = 0.5. If there are many ties (e.g., many values exactly at the hypothesized median), the sign test loses power and other methods (such as the Wilcoxon signed-rank test) may be preferred.

8.3.3 Normal Approximation for Large Samples

For $n \geq 20$, the binomial distribution is well approximated by a normal distribution. Using the continuity correction (adding or subtracting 0.5 to move from a discrete to a continuous distribution), the test statistic is:

$$z = (x^+ \pm 0.5 - n/2) / \sqrt{(n/4)}$$

where we use $x^+ - 0.5$ if $x^+ > n/2$, and $x^+ + 0.5$ if $x^+ < n/2$. The denominator $\sqrt{(n/4)} = \sqrt{n} / 2$ is the standard deviation of X^+. The resulting z is compared to the standard normal critical values.

8.3.4 One-Sided and Two-Sided Tests

The sign test is flexible for one-sided alternatives. For example, in many social science contexts, we want to test whether the median has increased (right-tailed) or decreased (left-tailed). The choice of one-tailed versus two-tailed should be made before examining the data, based on the research question.

Example 8.1 (Sign Test for Median Monthly Income — Right-tailed)

A sociologist in Lahore claims that the median monthly household income in a certain neighborhood exceeds Rs. 50,000. A random sample of 15 households is drawn, and their incomes (in thousand rupees) are: 48, 52, 55, 49, 58, 60, 47, 53, 51, 56, 54, 49, 52, 57, 59. Test the claim at the 5% level of significance.

Step 1. Hypotheses. Let η be the population median monthly income. The claim is $\eta > 50$, so we set: H_0: $\eta = 50$ versus H_1: $\eta > 50$ (right-tailed).

Step 2. $\alpha = 0.05$.

Step 3. Test statistic. Count the number of observations greater than 50. Values > 50: 52, 55, 58, 60, 53, 51, 56, 54, 52, 57, 59 → that is 11 observations. (Values equal to 50? None.) So $x^+ = 11$, n = 15. Under H_0, $X^+ \sim$ Binomial(15, 0.5).

Step 4. Critical region (p-value approach). Since this is a right-tailed test, the p-value = $P(X^+ \geq 11)$.

Step 5. Computation.

$P(X^+ \geq 11) = P(X^+=11) + P(X^+=12) + P(X^+=13) + P(X^+=14) + P(X^+=15)$

$= [C(15,11)(0.5)^{15}] + [C(15,12)(0.5)^{15}] + [C(15,13)(0.5)^{15}] + [C(15,14)(0.5)^{15}] + [C(15,15)(0.5)^{15}]$

$= (1365 + 455 + 105 + 15 + 1)/32768 = 1941/32768 \approx 0.0592.$

Step 6. Decision and conclusion. Since $p = 0.0592 > 0.05$, we fail to reject H_0. There is insufficient evidence at the 5% level to conclude that the median monthly income exceeds Rs. 50,000. (The result is borderline — a larger sample might yield significance.)

Example 8.2 (Sign Test for Median — Two-tailed, with a tie)

A nutritionist at a hospital in Karachi believes that the median body mass index (BMI) of adult patients is 22.5. A random sample of 12 patients gives BMIs: 21.2, 23.1, 22.5, 24.0, 20.8, 22.5, 23.5, 21.9, 22.0, 23.8, 22.5, 21.5. Test the hypothesis at $\alpha = 0.05$.

Step 1. H_0: $\eta = 22.5$ versus H_1: $\eta \neq 22.5$ (two-tailed).

Step 2. $\alpha = 0.05$.

Step 3. Discard observations exactly equal to 22.5. There are three such values (patients 3, 6, 11). After discarding, $n = 9$. Count values greater than 22.5: 23.1, 24.0, 23.5, 23.8 $\rightarrow x^+ = 4$.

Step 4. Under H_0, $X^+ \sim \text{Binomial}(9, 0.5)$. The two-tailed p-value $= 2 \times \min(P(X^+ \leq 4), P(X^+ \geq 4))$. We compute directly:

$P(X^+ \leq 4) = \Sigma_{k=0}^{4} C(9,k)(0.5)^9 = (1+9+36+84+126)/512 = 256/512 = 0.5$ exactly.

Therefore the two-tailed p-value $= 2 \times 0.5 = 1.0$. There is no evidence against H_0.

Step 5. Fail to reject H_0. The data are consistent with a median of 22.5.

8.4 THE SIGN TEST FOR PAIRED DATA

The sign test can be extended to paired data, such as before-after measurements or matched pairs. For each pair, we compute the difference $d_i = x_{1i} - x_{2i}$. Under the null hypothesis that there is no systematic difference between the two conditions, the median difference is zero. Therefore, we apply the one-sample sign test to the differences, testing H_0: median difference = 0.

The procedure is identical to the one-sample sign test: count the number of positive differences, discard zeros, and compare to the binomial distribution with $p = 0.5$.

Example 8.3 (Sign Test for Paired Data — Before/After Health Campaign)

A public health campaign in Karachi aims to reduce the number of sugary drinks consumed per week by teenagers. A random sample of 12 teenagers reports their weekly consumption before and after the campaign. The data (number of drinks) are:

Teenager	1	2	3	4	5	6	7	8	9	10	11	12
Before	8	6	10	5	7	9	8	6	7	10	5	8
After	6	5	8	4	6	7	6	5	6	8	4	7

Test whether the campaign reduced consumption (i.e., whether the median difference is positive) at $\alpha = 0.05$.

Step 1. Let d = Before − After. Positive differences indicate a reduction. H_0: median difference = 0 versus H_1: median difference > 0.

Step 2. $\alpha = 0.05$.

Step 3. Compute differences: (2, 1, 2, 1, 1, 2, 2, 1, 1, 2, 1, 1). All differences are positive, so $x^+ = 12$. No ties. n = 12.

Step 4. p-value = $P(X^+ \geq 12) = P(X^+ = 12) = (0.5)^{12} = 1/4096 \approx 0.000244$.

Step 5. Since $p < 0.05$, reject H_0.

Step 6. There is very strong evidence that the median consumption decreased after the campaign.

Example 8.4 (Sign Test for Paired Data — One-sided, with a zero difference)

A psychologist in Islamabad tests whether a new relaxation technique reduces anxiety scores. Ten patients are measured before and after the technique. Differences (before − after) are: 3, 5, 0, 2, 4, 1, 6, 0, 3, 2. Test at $\alpha = 0.05$ whether the technique reduces anxiety (i.e., median difference > 0).

Solution. Discard the two zeros (patients 3 and 8). Then n = 8. Positive differences: all remaining 8 differences are positive (3, 5, 2, 4, 1, 6, 3, 2 → all > 0), so $x^+ = 8$.

$p = P(X^+ \geq 8) = P(X^+ = 8) = (0.5)^8 = 1/256 \approx 0.0039$.

Reject H_0. There is strong evidence that the technique reduces anxiety.

8.5 THE WILCOXON SIGNED-RANK TEST FOR PAIRED DATA

The sign test uses only the direction of the differences, not their magnitude. This is a waste of information, because a large positive difference (say, 10 units) provides stronger evidence against the null than a small positive difference (say, 0.1 units), yet the sign test treats them identically. The Wilcoxon signed-rank test (developed by Frank Wilcoxon in 1945) remedies this by also considering the relative sizes of the differences. It is the nonparametric counterpart of the paired t-test and is substantially more powerful than the sign test when the differences are approximately symmetric.

8.5.1 Theoretical Foundation: Why Ranks?

The Wilcoxon signed-rank test is based on the following idea: if the differences are independent and come from a continuous symmetric distribution (not necessarily normal), then the absolute values of the differences are independent of their signs. Under the null hypothesis that the median difference is zero, the signs are equally likely to be positive or negative, and the ranks of the absolute values are fixed. Therefore, the sum of the ranks of the positive differences, W^+, has a known distribution that can be derived combinatorially.

The test is more powerful than the sign test because it gives higher weight to larger absolute differences. A large positive difference contributes a large rank to W^+, making the test more sensitive to deviations from the null.

8.5.2 Exact Distribution of W^+

For a sample of n non-zero differences, the possible values of W^+ range from 0 (all differences negative) to $n(n+1)/2$ (all differences positive). Under H_0, each assignment of signs to the ranks is equally likely, with probability $1/2^n$. Therefore, the distribution of W^+ is symmetric about its mean $\mu = n(n+1)/4$. For small n, exact critical values are tabulated. For example, for n = 5, the distribution of W^+ is:

W^+	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Prob.	1/32	1/32	2/32	2/32	3/32	3/32	4/32	4/32	4/32	4/32	3/32	3/32	2/32	2/32	1/32	1/32

The mean is 5×6/4 = 7.5, which matches the symmetry.

8.5.3 Mean and Variance under H_0

For any n, the mean and variance of W^+ are:

$$\mu_{W^+} = n(n+1)/4, \quad \sigma^2_{W^+} = n(n+1)(2n+1)/24$$

These are derived from the fact that the ranks are fixed numbers 1, 2, ..., n, and under H_0 the signs are independent Bernoulli(0.5). Hence, $W^+ = \Sigma I_i \times r_i$, where I_i are independent indicators for positive sign. Then $E[W^+] = \Sigma r_i \times 0.5 = (n(n+1)/2) \times 0.5 = n(n+1)/4$. The variance is $\Sigma r_i^2 \times Var(I_i) = \Sigma r_i^2 \times 0.25 = (1/4) \times n(n+1)(2n+1)/6 = n(n+1)(2n+1)/24$.

8.5.4 Normal Approximation and Continuity Correction

For $n \geq 15$, the distribution of W^+ is approximately normal. We use:

$$z = (W^+ - \mu_{W^+}) / \sigma_{W^+}$$

with a continuity correction of ±0.5. For a two-tailed test:

$$z = (W^+ - 0.5 - \mu_{W^+}) / \sigma_{W^+} \; if \; W^+ > \mu, \qquad or \; with \; +0.5 \; if \; W^+ < \mu$$

8.5.5 Ties and Their Effect

If there are ties among the absolute differences, we assign average ranks. This reduces the variance slightly, and the normal approximation can be adjusted using a correction factor. For simplicity, most statistical software automatically applies the correction. If there are many ties, the Wilcoxon test loses some power, but it remains valid. If there are ties within pairs (i.e., differences of zero), those pairs are discarded, as in the sign test.

Example 8.5 (Wilcoxon Signed-Rank Test — Before/After Blood Pressure)

A doctor in Islamabad tests a new relaxation technique on 10 patients with mild hypertension. Systolic blood pressure (mmHg) is recorded before and after a 4-week program. The data are:

Patient	Before	After	Difference (Before − After)
1	142	138	4
2	150	145	5
3	136	132	4
4	148	140	8
5	155	150	5
6	140	138	2
7	144	142	2
8	152	146	6
9	138	135	3
10	146	144	2

Test whether the technique reduces blood pressure (i.e., median difference > 0) at $\alpha = 0.05$.

Step 1. H_0: median difference = 0 versus H_1: median difference > 0.

Step 2. $\alpha = 0.05$.

Step 3. Compute absolute differences and ranks (excluding zeros — none here):

d	|d|	Rank	Signed rank
4	4	4.5 (tie)	+4.5
5	5	7.5 (tie)	+7.5
4	4	4.5	+4.5
8	8	10	+10
5	5	7.5	+7.5
2	2	2 (tie)	+2
2	2	2	+2
6	6	9	+9
3	3	3	+3
2	2	2	+2

All differences are positive, so $W^+ = 4.5+7.5+4.5+10+7.5+2+2+9+3+2 = 52.0$.

Step 4. Under H_0, $\mu = n(n+1)/4 = 10\times11/4 = 27.5$. $\sigma = \sqrt{[10\times11\times21/24]} = \sqrt{(2310/24)} = \sqrt{96.25} = 9.81$.

Step 5. $z = (52 - 27.5)/9.81 = 24.5/9.81 = 2.50$. The one-tailed p-value ≈ 0.0062 (from normal table).

Step 6. Since $p < 0.05$, reject H_0. There is strong evidence that the relaxation technique reduces systolic blood pressure.

Example 8.6 (Wilcoxon Signed-Rank Test — Two-tailed, with ties)

A researcher in Faisalabad compares two different study methods for learning a foreign language. Ten students use both methods (order randomized). The difference in test scores (Method A − Method B) are: 2, −1, 3, 0, −2, 4, 1, −3, 2, −1. Test whether the methods differ at $\alpha = 0.05$.

Step 1. H_0: median difference = 0 versus H_1: median difference ≠ 0.

Step 2. $\alpha = 0.05$.

Step 3. Discard the zero difference (one student). n = 9. Absolute differences: 2,1,3,2,4,1,3,2,1. Sort: 1,1,1,2,2,2,3,3,4. Average ranks: 1 → rank 2; 2 → rank 5; 3 → rank 7.5; 4 → rank 9.

Positive differences (d>0): 2, 3, 4, 2, 2 → ranks: 5, 7.5, 9, 5, 5 → sum = 31.5 = W^+.

Step 4. $\mu = 9\times10/4 = 22.5$. $\sigma = \sqrt{[9\times10\times19/24]} = \sqrt{71.25} = 8.44$.

Step 5. $z = (31.5 - 22.5)/8.44 = 9/8.44 = 1.067$. Two-tailed $p = 2\times P(Z>1.067) \approx 2\times0.143 = 0.286$.

Step 6. Fail to reject H_0. No evidence that the methods differ.

8.6 THE MANN-WHITNEY U TEST (WILCOXON RANK-SUM TEST)

The Mann-Whitney U test (also called the Wilcoxon rank-sum test) is the nonparametric alternative to the independent-samples t-test. It tests whether two independent samples come from populations with the same distribution, against the alternative that one population tends to produce larger values than the other. The test was developed independently by Frank Wilcoxon (1945) and by Henry Mann and Donald Whitney (1947). It is one of the most widely used nonparametric tests in the social sciences.

8.6.1 Conceptual Basis: Stochastic Superiority

The Mann-Whitney test does not test equality of means (unless additional assumptions hold). Instead, it tests a more general hypothesis: stochastic equality versus stochastic superiority. Formally, let X be a random observation from population 1 and Y from population 2. The null hypothesis is:

$$H_0: P(X > Y) = P(X < Y) = 0.5$$

That is, the probability that a randomly chosen X exceeds a randomly chosen Y is the same as the probability that Y exceeds X. Under the alternative (one-sided), we have $P(X > Y) > 0.5$ (X tends to be larger) or $P(X > Y) < 0.5$ (Y tends to be larger). If the two distributions have the same shape (i.e., they are identical except possibly for a shift), then this is equivalent to testing equality of medians.

8.6.2 The U Statistic and Its Relationship to W

The Mann-Whitney U statistic is defined as the number of times an observation from sample 1 precedes an observation from sample 2 when all observations are sorted together. Equivalently, if we rank all $N = n_1 + n_2$ observations and let R_1 be the sum of the ranks of sample 1, then:

$$U = R_1 - n_1(n_1+1)/2$$

U ranges from 0 to n_1n_2. Under H_0, the distribution of U is symmetric about $n_1n_2/2$. The test is often performed using the smaller of U and its complement $U' = n_1n_2 - U$. For one-sided tests, we use U directly.

8.6.3 Mean and Variance under H_0

Under the null hypothesis that the two distributions are identical, the ranks 1,...,N are randomly assigned to the two samples. Therefore, the expected rank sum for sample 1 is:

$$\mu_{\{R_1\}} = n_1 \cdot (N+1)/2$$

and the variance is:

$$\sigma^2_{\{R_1\}} = n_1n_2(N+1)/12$$

Since $U = R_1 - n_1(n_1+1)/2$, it follows that:

$$\mu_U = n_1n_2/2 , \quad \sigma^2_U = n_1n_2(N+1)/12$$

These formulas assume no ties. When ties are present, the variance is adjusted downward.

8.6.4 Normal Approximation

For $n_1, n_2 \geq 8$, the distribution of U is approximately normal. The test statistic is:

$$z = (U - \mu_U) / \sigma_U$$

with a continuity correction (add or subtract 0.5 to U). For a two-tailed test:

$$z = (U - n_1n_2/2 \pm 0.5) / \sqrt{[n_1n_2(N+1)/12]}$$

8.6.5 Handling Ties in the Rank-Sum Test

When there are ties, we assign average ranks. This reduces the variance. The exact formula for the variance under ties is more complicated, but most software computes it automatically. For hand calculation, if ties are few, the uncorrected variance is a slight overestimate, making the test conservative (p-value slightly larger than true). For a conservative test, we can ignore the tie correction; for a more accurate test, we use software.

Example 8.7 (Mann-Whitney U Test — Comparing Two Teaching Methods)

An educational researcher in Multan wants to compare two methods of teaching mathematics to 8th-grade students. Method A (innovative) is used in one school, Method B (traditional) in another. A random sample of 8 students from each school takes a standardized test. Scores (out of 100) are:

Method A: 78, 85, 72, 90, 68, 88, 75, 82

Method B: 70, 74, 65, 80, 68, 72, 66, 69

Step 1. H_0: The two distributions are identical. H_1: They are not identical (one tends to be larger).

Step 2. $\alpha = 0.05$.

Step 3. Combine and rank all 16 scores (ties at 68 and 72 each get average ranks):

Score	Group	Rank
65	B	1
66	B	2
68	A	3.5
68	B	3.5
69	B	5
70	B	6
72	A	7.5
72	B	7.5
74	B	9
75	A	10
78	A	11
80	B	12
82	A	13
85	A	14
88	A	15
90	A	16

Sum of ranks for Method A (n_1=8): 3.5+7.5+10+11+13+14+15+16 = 90.0.

$U = R_1 - n_1(n_1+1)/2 = 90 - 8\times9/2 = 90 - 36 = 54$.

Step 4. $\mu = n_1n_2/2 = 8\times8/2 = 32$. $\sigma = \sqrt{[8\times8\times17/12]} = \sqrt{90.667} = 9.52$.

Step 5. $z = (54 - 32)/9.52 = 22/9.52 = 2.31$. Two-tailed p-value = $2\times P(Z>2.31) \approx 2\times0.0104 = 0.0208$.

Step 6. Since $p < 0.05$, reject H_0. There is evidence that the distributions differ; Method A appears to produce higher scores.

Example 8.8 (Mann-Whitney U Test — One-tailed, with unequal sample sizes)

A sociologist in Peshawar believes that female university students spend more time on social media than male students. A random sample of 12 female and 10 male students yields weekly hours on social media (self-reported).

Females: 14, 18, 12, 20, 15, 22, 16, 19, 17, 21, 13, 24

Males: 10, 12, 8, 15, 11, 13, 9, 14, 10, 12

Step 1. H_0: Distributions identical. H_1: Female hours are stochastically larger (right-tailed).

Step 2. $\alpha = 0.05$.

Step 3. Rank all 22 observations. The sum of ranks for females ($n_1=12$) is $R_1 = 168.5$. Then $U = 168.5 - 12\times13/2 = 168.5 - 78 = 90.5$.

Step 4. $\mu = n_1n_2/2 = 12\times10/2 = 60$. $\sigma = \sqrt{[12\times10\times23/12]} = \sqrt{230} = 15.17$.

Step 5. $z = (90.5 - 60)/15.17 = 30.5/15.17 = 2.01$. One-tailed $p = P(Z>2.01) = 0.0222$.

Step 6. Since $p < 0.05$, reject H_0. The data support the claim that female students spend significantly more hours on social media.

8.7 THE KRUSKAL-WALLIS H TEST FOR SEVERAL INDEPENDENT SAMPLES

The Kruskal-Wallis H test is the nonparametric equivalent of one-way ANOVA. It compares three or more independent groups to determine whether at least one group tends to produce larger values than the others. The test was developed by William Kruskal and W. Allen Wallis in 1952. It is an extension of the Mann-Whitney U test to k > 2 groups.

8.7.1 Extension of the Mann-Whitney to k > 2

The logic of the Kruskal-Wallis test is straightforward: combine all observations from the k groups, rank them from 1 to N, and then compute the sum of ranks for each group. If the groups are identical, the rank sums should be roughly proportional to the sample sizes. If one group tends to have larger values, its rank sum will be larger than expected. The test statistic H measures the discrepancy between the observed rank sums and the expected rank sums under the null hypothesis.

8.7.2 The H Statistic and Its Chi-Square Approximation

The Kruskal-Wallis H statistic is given by:

$$H = [12 / N(N+1)] \times \Sigma (R_i^2/n_i) - 3(N+1)$$

where k = number of groups, n_i = sample size of group i, N = total sample size = Σn_i, and R_i = sum of ranks for group i. Under the null hypothesis that all k populations have the same distribution, and for sufficiently large n_i (usually each $n_i \geq 5$), H follows a chi-square distribution with k − 1 degrees of freedom. The test is always one-tailed (upper-tail), because H increases as the rank sums become more spread out.

8.7.3 Post-Hoc Comparisons after a Significant H

If H is significant (i.e., $H > \chi^2\alpha, k - 1$), we reject the null hypothesis that all groups are equal. However, we do not yet know which groups differ. To identify specific pairwise differences, we can perform Mann-Whitney U tests between each pair of groups, but we must adjust the significance level to control the family-wise Type I error rate. The Bonferroni correction is simple: if we perform m = k(k−1)/2 pairwise tests, we use $\alpha' = \alpha/m$. For example, with k = 4, m = 6, and $\alpha = 0.05$, we would test each pair at $\alpha' = 0.05/6 \approx 0.0083$.

8.7.4 Assumptions and Robustness

The Kruskal-Wallis test assumes: independent random samples from each population; the response variable is continuous (or at least ordinal); and the distributions have the same shape if we want to interpret the test as comparing medians. If shapes differ, the test still detects stochastic differences but the interpretation is more general. The test is robust to outliers and does not require normality. It is about 95% as efficient as one-way ANOVA when the normality assumption holds, and can be much more efficient when normality is violated.

Example 8.9 (Kruskal-Wallis Test — Comparing Three Fertilizer Types)

An agricultural experiment in Faisalabad compares the yield (kg per plot) of wheat under three fertilizer treatments. The yields are:

Fertilizer X: 42, 45, 48, 44, 46

Fertilizer Y: 38, 40, 42, 39, 41

Fertilizer Z: 50, 52, 49, 51, 53

Test at $\alpha = 0.05$ whether the three treatments differ.

Step 1. H_0: All three distributions are identical. H_1: At least one differs.

Step 2. $\alpha = 0.05$.

Step 3. Combine and rank the 15 observations (tie at 42: ranks 5 and 6 → average 5.5):

Value	Group	Rank
38	Y	1
39	Y	2
40	Y	3
41	Y	4
42	X	5.5
42	Y	5.5
44	X	7
45	X	8
46	X	9
48	X	10
49	Z	11
50	Z	12
51	Z	13
52	Z	14
53	Z	15

Rank sums: $R_x = 5.5+7+8+9+10 = 39.5$; $R_y = 1+2+3+4+5.5 = 15.5$; $Rz = 11+12+13+14+15 = 65$.

Step 4. N=15, N(N+1)=240.

$\Sigma(R_i^2/n_i) = (39.5^2/5)+(15.5^2/5)+(65^2/5) = 312.05+48.05+845 = 1205.1.$

$H = (12/240)\times 1205.1 - 3\times 16 = 0.05\times 1205.1 - 48 = 60.255 - 48 = 12.255.$

Step 5. df = k−1 = 2. Critical $\chi^2_{0.05,2} = 5.991$. Since H = 12.255 > 5.991, reject H_0.

Step 6. There is strong evidence that the fertilizers produce different yields. Post-hoc Mann-Whitney tests (with Bonferroni correction) would show that Fertilizer Z yields significantly higher than both X and Y.

Example 8.10 (Kruskal-Wallis Test — Four Teaching Methods)

A university in Lahore compares four different teaching methods for a statistics course. Twenty students are randomly assigned to the four methods (5 per method). Final exam scores are:

Method 1: 65, 70, 68, 72, 66

Method 2: 75, 78, 72, 80, 76

Method 3: 60, 62, 58, 65, 61

Method 4: 85, 82, 88, 84, 86

Step 1. H_0: All four distributions identical. H_1: At least one differs.

Step 2. $\alpha = 0.05$.

Step 3. All 20 values sorted; ties at 65 (M1 and M3) get rank 5.5; ties at 72 (M1 and M2) get rank 10.5. Rank sums:

R_1 (M1) = 5.5+7+8+9+10.5 = 40; R_2 (M2) = 10.5+12+13+14+15 = 64.5

R_3 (M3) = 1+2+3+4+5.5 = 15.5; R_4 (M4) = 16+17+18+19+20 = 90

Step 4. N=20, N(N+1)=420.

$\Sigma(R_i^2/n_i) = (40^2/5)+(64.5^2/5)+(15.5^2/5)+(90^2/5) = 320+832.05+48.05+1620 = 2820.1.$

$H = (12/420)\times 2820.1 - 3\times 21 = 80.57 - 63 = 17.57.$

Step 5. df = 3, critical $\chi^2_{0.05,3} = 7.815$. Since 17.57 > 7.815, reject H_0.

Step 6. There is strong evidence that the teaching methods differ. Method 4 appears best, Method 3 worst.

8.8 THE ONE-SAMPLE RUNS TEST FOR RANDOMNESS

The runs test is used to determine whether a sequence of binary outcomes (or a dichotomised sequence) is random. It is often applied to check the assumption of independence in time series or to test whether the order of observations in a sample is random. The test was developed by Swed and Eisenhart (1943). A run is a maximal consecutive sequence of identical outcomes. Too few runs indicate clustering (positive autocorrelation), while too many runs indicate alternation (negative autocorrelation).

8.8.1 The Concept of a Run

Consider a sequence of two symbols, say H (heads) and T (tails). For example: H H T T H T T T. This sequence has runs: HH (run 1), TT (run 2), H (run 3), TTT (run 4). So R = 4. If the sequence were perfectly alternating (H T H T H T ...), the number of runs would be maximized. If it were all H followed by all T, there would be only 2 runs.

8.8.2 Exact Distribution of the Number of Runs

Under the null hypothesis that the sequence is random (i.e., all orders of n_1 symbols of type 1 and n_2 symbols of type 2 are equally likely), the exact distribution of the number of runs R is known. For small n_1 and n_2, exact critical values are tabulated.

8.8.3 Mean and Variance under Randomness

For large n_1, n_2 (both ≥ 10), the distribution of R is approximately normal with:

$$\mu_R = 2n_1n_2/(n_1 + n_2) + 1$$
$$\sigma^2{}_R = 2n_1n_2(2n_1n_2 - n_1 - n_2) / [(n_1 + n_2)^2(n_1 + n_2 - 1)]$$

8.8.4 Normal Approximation

The test statistic is:

$$z = (R - \mu_R) / \sigma_R$$

with a continuity correction of ±0.5. A two-tailed test rejects if $|z| > z_{\alpha/2}$.

8.8.5 Using the Runs Test with Continuous Data (Above/Below Median)

For continuous data, we can convert the sequence to a binary sequence by comparing each observation to the sample median. Values above the median become "A" (or 1), values below become "B" (or 0). Observations exactly equal to the median are discarded. Then we perform the runs test on the resulting binary sequence.

Example 8.11 (Runs Test — Quality Control)

A factory in Sialkot records whether each produced item is defective (D) or non-defective (N). The sequence of 30 items is:

N N N D N N D D N N N N D N N N D D D N N N N D N N N N D D

Solution.

Count runs: NNN (run 1), D (run 2), NN (run 3), DD (run 4), NNNN (run 5), D (run 6), NNN (run 7), DDD (run 8), NNNN (run 9), D (run 10), NNNN (run 11), DD (run 12). So R = 12.

$n_1 = 20$ (N), $n_2 = 10$ (D).

$\mu^R = (2\times20\times10)/30 + 1 = 400/30 + 1 = 14.333$.

$\sigma^{2R} = [2\times20\times10\times(400-30)] / (30^2\times29) = (400\times370)/26100 = 148000/26100 = 5.670$, so $\sigma^R = 2.381$.

$z = (12 - 14.333)/2.381 = -2.333/2.381 = -0.98$. $|z| = 0.98 < 1.96$, fail to reject H_0.

Conclusion: The sequence is consistent with randomness.

Example 8.12 (Runs Test — Above/Below Median for Temperature Data)

The average summer temperatures (°F) for 20 successive years from 1971 to 1990 in a given city are: 72, 71, 70, 82, 80, 77, 71, 85, 75, 80, 82, 81, 83, 82, 85, 86, 83, 81, 82, 84. Test whether the data constitute a random sample.

Solution.

Sorted values: 70,71,71,72,75,77,80,80,81,81,82,82,82,82,83,83,84,85,85,86. The median = (81+82)/2 = 81.5.

Binary sequence (1 if >81.5, 0 if ≤81.5):

0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1

Count runs: 000 (1), 1 (2), 000 (3), 1 (4), 00 (5), 1 (6), 0 (7), 1 (8), 11111 (9), 0 (10), 11 (11). So R = 11.

$n_1 = 10$ (ones), $n_2 = 10$ (zeros). $\mu^R = (2\times10\times10)/20 + 1 = 11$. $\sigma^{2R} = [200\times(200-20)]/(400\times19) = 200\times180/7600 = 4.737$, $\sigma^R = 2.176$.

$z = (11 - 11)/2.176 = 0$. $p = 1.0$. Fail to reject randomness.

8.9 THE SPEARMAN RANK CORRELATION TEST

The Spearman rank correlation coefficient (ρ or r_s) is a nonparametric measure of the strength and direction of the monotonic relationship between two variables. It is computed by applying the Pearson correlation formula to the ranks of the data rather than to the raw values. It is the nonparametric alternative to Pearson's r and is widely used when the relationship is monotonic but not necessarily linear, or when the data contain outliers.

8.9.1 From Pearson to Spearman: The Rank Transformation

Pearson's correlation measures linear association. If the relationship is curved (e.g., exponential), Pearson's r may be close to zero even though the variables are strongly related. Spearman's ρ, by converting to ranks, measures monotonic association: as X increases, does Y tend to increase (or decrease) consistently? The rank transformation is:

- Replace each X_i by its rank among all X's: $R(X_i)$
- Replace each Y_i by its rank among all Y's: $R(Y_i)$
- Compute Pearson's r on the pairs of ranks.

8.9.2 Derivation of the Simplified Formula

If there are no ties, the simplified formula is:

$$r_s = 1 - 6\Sigma d_i^2 / [n(n^2 - 1)]$$

where $d_i = R(X_i) - R(Y_i)$. This formula is derived from the fact that the sum of squared ranks is fixed. When ties exist, we must use the full Pearson formula on the ranks.

8.9.3 Ties and the Exact Calculation

When ties occur, we assign average ranks. Then we compute the Pearson correlation of the ranks:

$$r_s = \Sigma(R(X_i) - R_x)(R(Y_i) - R_y) / \sqrt{[\Sigma(R(X_i) - R_x)^2 \, \Sigma(R(Y_i) - R_y)^2]}$$

8.9.4 Hypothesis Test for ρ_s

To test H_0: $\rho_s = 0$ versus H_1: $\rho_s \neq 0$ (or one-sided), for $n \geq 10$ we can use the t-test:

$$t = r_s \sqrt{[(n-2)/(1-r_s^2)]}$$

which follows a t-distribution with $n - 2$ degrees of freedom under H_0. For small n, exact tables are available.

8.9.5 Relationship to Kendall's Tau

Another common nonparametric correlation measure is Kendall's τ (tau). While Spearman's ρ is based on squared rank differences, Kendall's τ is based on the number of concordant and discordant pairs. Both are valid; Spearman is more widely used in social science research.

Example 8.13 (Spearman Rank Correlation — Study Time and Exam Rank)

Ten students in a BS Economics class at AIOU are ranked by hours studied per week (X) and by final exam score (Y). The ranks (already in order) are:

Student	Rank X	Rank Y	d = X−Y	d^2
1	1	2	−1	1
2	2	1	1	1
3	3	4	−1	1
4	4	3	1	1
5	5	5	0	0
6	6	7	−1	1
7	7	6	1	1
8	8	9	−1	1
9	9	8	1	1
10	10	10	0	0

$\Sigma d^2 = 8$.

$r_s = 1 - (6\times8)/(10\times(100-1)) = 1 - 48/990 = 1 - 0.0485 = 0.9515$.

Test H_0: $\rho_s = 0$. $t = 0.9515\times\sqrt{(8/0.095)} = 0.9515\times\sqrt{84.21} = 0.9515\times9.177 = 8.73$. With df=8, $p < 0.001$. Strong evidence of a positive monotonic relationship.

Example 8.14 (Spearman Correlation — Outlier Resistance)

Consider two variables: X = 1,2,3,4,5,100 and Y = 2,4,6,8,10,200. Pearson's r = 1.0 (perfect linear). Now add an outlier: X = 1,2,3,4,5,100, and Y = 2,4,6,8,10,50 (the last point is changed to 50 instead of 200). Pearson's r becomes about 0.78, while Spearman's r_s remains 0.94 because the ranks are unchanged (the largest X still pairs with the largest Y in rank order). This illustrates robustness to outliers.

8.10 THE KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

The Kolmogorov-Smirnov (K-S) test is a nonparametric test for comparing an empirical distribution with a theoretical distribution (one-sample) or for comparing two empirical distributions (two-sample). Unlike the chi-square goodness-of-fit test, which requires arbitrary binning of continuous data, the K-S test works directly with the raw data and is sensitive to differences in location, shape, and spread.

8.10.1 The Empirical Distribution Function (ECDF)

For a sample of n observations $x_1,...,x_n$, the empirical cumulative distribution function (ECDF) is defined as:

$$F_n(x) = (number\ of\ observations \leq x) / n$$

It is a step function that increases by 1/n at each observed value.

8.10.2 The One-Sample K-S Statistic

The one-sample K-S test statistic is the maximum absolute difference between the ECDF and the theoretical CDF $F_0(x)$:

$$D_n = sup_x |F_n(x) - F_0(x)|$$

Under H_0 that the sample comes from F_0, the distribution of D_n is known and tabulated. For large n, the critical value at $\alpha = 0.05$ is approximately $1.36/\sqrt{n}$, and at $\alpha = 0.01$ it is approximately $1.63/\sqrt{n}$.

8.10.3 Critical Values and the Lilliefors Correction

If the parameters of the theoretical distribution are estimated from the sample (e.g., estimating μ and σ for a normal distribution), the standard K-S critical values are too conservative (they give p-values that are too large). The Lilliefors test provides corrected critical values for normality with estimated parameters. Most statistical software includes the Lilliefors correction automatically.

8.10.4 The Two-Sample K-S Test

For two independent samples with ECDFs $F_n(x)$ and $G_m(x)$, the two-sample K-S statistic is:

$$D_{n,m} = sup_x |F_n(x) - G_m(x)|$$

The null hypothesis is that the two samples come from the same continuous distribution. For large samples, the approximate critical value at $\alpha = 0.05$ is $1.36\sqrt{[(n+m)/(nm)]}$.

Example 8.15 (K-S Test — Normality of Examination Scores)

A teacher in Peshawar wants to check whether 20 examination scores are normally distributed with mean 70 and standard deviation 10. The K-S test yields $D_n = 0.12$. With n=20, critical value at α=0.05 is $1.36/\sqrt{20} = 0.304$. Since $0.12 < 0.304$, do not reject normality.

Example 8.16 (Two-Sample K-S Test — Comparing Two Distributions)

A researcher in Karachi compares the age distributions of two groups of workers (n_1=30, n_2=35). The K-S statistic D = 0.22. The critical value at α=0.05 is $1.36\sqrt{[(30+35)/(30\times35)]} = 1.36\times0.249 = 0.339$. Since $0.22 < 0.339$, fail to reject H_0. No evidence that the age distributions differ.

8.11 FACTORS AFFECTING THE MAGNITUDE OF CORRELATION

In both parametric (Pearson) and nonparametric (Spearman) correlation, the magnitude of the coefficient can be influenced by several factors. Understanding these is crucial for correct interpretation.

8.11.1 Range restriction

If the range of one variable is artificially limited (e.g., studying only students with high GPAs), the correlation tends to shrink. This is because variance is reduced, limiting the possible covariance.

8.11.2 Outliers

A single outlier can dramatically inflate or deflate Pearson's r, but Spearman's r_s is much less affected because ranks are not sensitive to extreme magnitudes.

8.11.3 Non-linearity

Pearson's r may be near zero even when a strong non-linear relationship exists (e.g., a U-shaped curve). Spearman's r_s captures monotonic relationships but may also miss non-monotonic patterns. Always plot the data first.

8.11.4 Heteroscedasticity

Unequal spread across the range can distort Pearson's r. Spearman is robust to this.

8.11.5 Measurement error

Random measurement error attenuates both correlation coefficients, pulling them toward zero. The more noise, the weaker the observed correlation.

8.11.6 Sample size

Very small samples produce unstable estimates; larger samples give more precise estimates. A correlation of 0.3 with n=10 is not significant; with n=100 it may be highly significant.

8.12 COMPARING PARAMETRIC AND NONPARAMETRIC TESTS: STRENGTHS AND WEAKNESSES

We now summarize the key differences between parametric and nonparametric tests, focusing on asymptotic relative efficiency (ARE) and power under various conditions.

8.12.1 Asymptotic Relative Efficiency

The ARE of a nonparametric test relative to its parametric counterpart is the ratio of sample sizes needed to achieve the same power when the parametric assumptions hold. For the Wilcoxon signed-rank test versus the t-test, ARE = $3/\pi \approx 0.955$. For the Mann-Whitney versus the two-sample t-test, ARE = $3/\pi$ as well. For the Kruskal-Wallis versus ANOVA, ARE approaches 0.955 for large k. This means that if the t-test requires n = 100 to achieve a certain power, the Wilcoxon test requires about 105 observations to achieve the same power when the data are normal. The loss is very small.

When the data are not normal, the nonparametric test can have ARE much greater than 1. For heavy-tailed distributions (e.g., Cauchy), the ARE can be infinite, meaning the t-test completely fails while the Wilcoxon test works perfectly.

8.12.2 Power Comparisons under Different Distributions

- Normal distribution: t-test has slightly higher power (by about 5%).
- Light-tailed (uniform): t-test and Wilcoxon similar.
- Heavy-tailed (t with 3 df): Wilcoxon much more powerful.
- Skewed (log-normal): Wilcoxon more powerful, especially for small samples.
- Outliers present: Wilcoxon far superior.

8.12.3 Advantages of Nonparametric Tests

- Fewer assumptions.
- Valid for ordinal data.

- Robust to outliers.
- Exact p-values for small samples (permutation basis).
- Often as powerful as parametric tests even when assumptions hold.

8.12.4 Disadvantages

- Less familiar to some researchers.
- May be less efficient for very small samples when normality holds (but the difference is tiny).
- Confidence intervals are more difficult to construct (though possible).
- Some tests (e.g., Mann-Whitney) are often misinterpreted as tests of medians without checking shape assumptions.

8.13 CHOOSING THE RIGHT NONPARAMETRIC TEST: A DECISION MAP

8.13.1 How many samples?

- One sample → test median: Sign test (if very simple) or Wilcoxon signed-rank (if more powerful and differences are symmetric).
- Two samples (paired) → Sign test (crude) or Wilcoxon signed-rank (preferred, more powerful).
- Two samples (independent) → Mann-Whitney U.
- Three or more independent samples → Kruskal-Wallis H.
- Two or more samples (dependent/repeated measures) → Friedman test (not covered in this unit; see advanced texts).

8.13.2 Type of question?

- Randomness of a sequence → Runs test.
- Association between two variables → Spearman rank correlation (monotonic) or Kendall's tau.
- Goodness of fit to a distribution → Kolmogorov-Smirnov (continuous) or chi-square (categorical).
- Comparing two distributions (two-sample) → Kolmogorov-Smirnov (general) or Mann-Whitney (stochastic ordering).

8.13.3 Check assumptions:

- For Wilcoxon signed-rank: differences should be roughly symmetric (not required for sign test).
- For Mann-Whitney: if testing medians, assume equal shapes; otherwise interpret as stochastic superiority.
- For Kruskal-Wallis: if testing medians, assume equal shapes; otherwise interpret as stochastic differences.

8.14 COMMON PITFALLS IN NONPARAMETRIC TESTING

8.14.1 Using a nonparametric test when a parametric test is clearly appropriate

If the data are normal and sample sizes are large, the parametric test is more powerful. Don't use nonparametric "just because".]

8.14.2 Ignoring ties

Ties require average ranks; failure to do so biases the test. In the sign test, ties at the hypothesized median are discarded, which reduces power.

8.14.3 Misinterpreting the null hypothesis

Rejecting H_0 in a Mann-Whitney test does not necessarily mean the medians differ; it means one distribution is stochastically larger. To conclude about medians, we need the additional assumption that the distributions have the same shape.

8.14.4 Small sample sizes without exact p-values

Normal approximations can be poor when $n < 10$. Use exact tables or software.

8.14.5 Using the runs test on non-dichotomous data without a valid splitting rule

Using the runs test on non-dichotomous data without a valid splitting rule. Splitting at the median reduces power; if possible, use a natural binary classification.

8.14.6 Over-reliance on p-values without effect size

A significant Mann-Whitney test does not tell us how large the difference is. Report a measure of effect size, such as the probability of superiority ($P(X>Y) = U/(n_1 n_2)$).

8.14.7 Applying nonparametric tests to data with very small sample sizes (n < 5) without caution

Applying nonparametric tests to data with very small sample sizes ($n < 5$) without caution. The smallest possible p-value for a sign test with $n=4$ is $2\times(0.5)^4 = 0.125$, so significance at $\alpha=0.05$ is impossible. Always check the achievable significance level.

8.15 SUMMARY

This unit has introduced the most important nonparametric methods for situations where parametric assumptions (normality, equal variances) are violated. The key points are:

- Nonparametric tests are distribution-free; they rely on ranks or signs rather than raw data.
- The sign test is the simplest — it uses only the direction of differences. It tests the median and is applicable even with very small samples, but has low power.
- The Wilcoxon signed-rank test is more powerful than the sign test for paired data because it incorporates the magnitude of differences via ranks.
- The Mann-Whitney U test is the nonparametric alternative to the two-sample t-test. It tests whether one distribution is stochastically larger than another.
- The Kruskal-Wallis H test extends the Mann-Whitney to three or more groups. It is the nonparametric analogue of one-way ANOVA.
- The runs test checks randomness of a binary sequence. It is useful for checking independence in time-ordered data.
- Spearman rank correlation measures monotonic association and is robust to outliers. It is the nonparametric alternative to Pearson's r.
- The Kolmogorov-Smirnov test compares an empirical distribution to a theoretical one (or two empirical distributions) without binning, making it ideal for continuous data.

- Nonparametric tests are about 95% as efficient as their parametric counterparts when normality holds, but are much safer when normality is in doubt. Their asymptotic relative efficiency is $3/\pi \approx 0.955$ for the Wilcoxon and Mann-Whitney tests compared to t-tests.
- When choosing a nonparametric test, consider the number of samples, whether they are paired or independent, the type of question (median, distribution, randomness, association), and the presence of ties.
- Common pitfalls include misinterpreting the null hypothesis, ignoring ties, using normal approximations with small samples, and applying the runs test without a valid dichotomization.

In Unit 9 we shall study time-series analysis and forecasting. The student who has mastered the material of the present unit will find the regression-based forecasting models a natural extension of the ideas developed here.

8.16 SELF-ASSESSMENT EXERCISE

1. Under what conditions would you prefer a nonparametric test over a parametric test? Give three specific examples from social science research in Pakistan.
2. What is the main difference between the sign test and the Wilcoxon signed-rank test? Why is the latter generally more powerful?
3. State the null and alternative hypotheses for the Mann-Whitney U test in plain language. What does it mean to say that "one distribution is stochastically larger" than another? How does this differ from a test of medians?
4. Explain how the Kruskal-Wallis H statistic is related to one-way ANOVA on ranks. Why does H have approximately a chi-square distribution? What is the role of the factor $12/(N(N+1))$?
5. What is a "run" in the runs test? How would you interpret a very small number of runs versus a very large number of runs? Give an example of each.
6. Derive the mean and variance of the Wilcoxon signed-rank statistic under the null hypothesis. Explain the logic of ranking absolute differences and then summing the ranks of the positive differences.
7. Compare and contrast the Pearson correlation coefficient and the Spearman rank correlation coefficient. In what situations would Spearman be preferred? Illustrate with a numerical example where the two coefficients give substantially different values.
8. Derive the mean and variance for the sampling distribution of the Wilcoxon signed-rank statistic W^+ under the null hypothesis. Start from the fact that the ranks are fixed and the signs are independent Bernoulli(0.5). Then explain how the normal approximation with continuity correction is obtained. Apply this to a numerical example of your own choosing.
9. For the Mann-Whitney U test, derive the mean and variance under H_0. Show that $U = n_1 n_2 - U'$. Explain how the test can be used for one-sided alternatives. Provide a complete worked example with ties.
10. Discuss the Kruskal-Wallis test in detail. Derive the H statistic, explain its relationship to the chi-square distribution, and describe the post-hoc procedures (Bonferroni-corrected Mann-Whitney tests) that follow a significant H. Use a numerical example with three groups.

11. A researcher reports a p-value of 0.03 from a Mann-Whitney U test with sample sizes $n_1 = 5$, $n_2 = 6$. Would you trust this p-value if it was computed using a normal approximation? Why or why not? What would you recommend instead?
12. You have paired data with 12 differences, all positive except two that are negative but very close to zero. The Wilcoxon signed-rank test gives a p-value of 0.08, while the sign test gives 0.12. Which test is more appropriate and why? Which test is more powerful in this situation?
13. In a Kruskal-Wallis test with $k = 4$ groups and total $N = 40$, the computed $H = 8.5$. The critical chi-square value at $\alpha = 0.05$ is 7.815. What do you conclude? What would you do next? If you perform post-hoc Mann-Whitney tests, what significance level would you use for each comparison to maintain an overall $\alpha = 0.05$?
14. A runs test on a sequence of 50 observations split at the median yields $R = 22$, with $n_1 = 25$, $n_2 = 25$. The mean number of runs under randomness is 26. Compute the approximate z-statistic and decide whether the sequence is random at $\alpha = 0.05$. Would your conclusion change if you used a continuity correction? Explain.
15. You compute Spearman's $r_s = 0.45$ from $n = 15$ pairs. Test whether this is significantly different from zero at $\alpha = 0.05$ (two-tailed). What is the critical value from the t-distribution? What would you conclude if the sample size were $n = 30$ with the same r_s?
16. A sample of 12 households in Rawalpindi has monthly water bills (Rs.): 1200, 1350, 1100, 1450, 1300, 1250, 1400, 1150, 1500, 1280, 1320, 1180. Test the hypothesis that the median bill is Rs. 1250 against the alternative that it is greater, using the sign test at $\alpha = 0.05$. [Answer: $x^+ = 7$, $p = 0.387$, fail to reject]
17. For the same data as in problem 16, use the Wilcoxon signed-rank test (one-sample version) to test H_0: median = 1250 versus H_1: median > 1250. [Answer: $W^+ = 48$, $\mu = 39$, $\sigma \approx 12.0$, $z = 0.75$, $p \approx 0.227$, fail to reject]
18. Two groups of students were taught statistics using two different methods. Method X (n=8) scores: 72, 68, 85, 90, 76, 80, 74, 88. Method Y (n=9) scores: 65, 70, 62, 78, 68, 72, 66, 69, 71. Use the Mann-Whitney U test at $\alpha = 0.05$ (two-tailed) to determine if the methods differ. [Answer: $R_x = 102.5$, $U = 66.5$, $\mu = 36$, $\sigma = 10.39$, $z = 2.93$, $p = 0.0034$, reject]
19. Four fertilizer treatments were applied to plots of wheat. Yields (kg/plot): A: 25,28,26; B: 22,24,23; C: 30,32,29; D: 27,26,28. Use the Kruskal-Wallis test at $\alpha = 0.05$. [Answer: $H = 8.45$, df=3, critical 7.815, reject]
20. A sequence of 20 coin tosses yields: H H T H T T H H H T H T T H T H H T T H. Test for randomness using the runs test at $\alpha = 0.05$. [Answer: n_1=10(H), n_2=10(T), R=13, μ=11, $\sigma \approx 2.18$, z=0.92, p=0.36, fail to reject]
21. Compute Spearman's rank correlation for the following pairs of ranks (X,Y): (1,2),(2,1),(3,3),(4,5),(5,4),(6,7),(7,6),(8,8). [Answer: $\Sigma d^2 = 6$, $r_s = 1 - 36/(8\times63) = 0.9286$]
22. Test the significance of $r_s = 0.9286$ from problem 21 with n=8. [Answer: $t = 0.9286\sqrt{(6/0.137)} = 6.15$, df=6, $p < 0.001$, significant]
23. A Kolmogorov-Smirnov test for normality on n=30 gives $D_n = 0.18$. The critical value at α=0.05 is $1.36/\sqrt{30} = 0.248$. What is your conclusion? [Answer: D_n < critical, do not reject normality]
24. For the water bill data in problem 16, perform the sign test for the two-sided alternative H_1: median $\neq$ 1250. [Answer: x^+=7, n=12, $p = 2\times P(X\geq7) = 2\times0.387 = 0.774$, fail to reject]

25. A Mann-Whitney U test with $n_1=6$, $n_2=8$ yields U = 12. Find the approximate p-value using the normal approximation. [Answer: $\mu=24$, $\sigma=\sqrt{60}=7.75$, $z=(12-24)/7.75=-1.55$, two-tailed $p\approx0.121$]

26. For the runs test in problem 20, comment on the direction of the observed number of runs relative to the expected. [Answer: R=13 is slightly above the mean 11, indicating mild alternation, but not significant.]

27. A Spearman correlation is computed from n=25 pairs, $r_s = 0.30$. Test H_0: $\rho_s = 0$ at $\alpha = 0.05$. [Answer: $t = 0.30\sqrt{(23/0.91)} = 0.30\times5.03 = 1.51$, df=23, critical ≈2.07, fail to reject]

UNIT 9: TIME SERIES & FORECASTING

KEY LEARNING OBJECTIVES

After studying this unit carefully and working through every solved example, the student should be able to:

1. Define a time series, list the four main components (secular trend, cyclical variation, seasonal variation and irregular variation), and explain the difference between the additive and multiplicative decomposition models.
2. Measure the secular trend of a time series by the freehand method, the method of semi-averages, the method of moving averages, and the method of least squares, and explain the strengths and weaknesses of each method.
3. Measure the seasonal variations in a time series by the ratio-to-trend method and the ratio-to-moving-average method, and use the resulting seasonal indices to deseasonalize the data.
4. Estimate the cyclical component of a time series by the residual method and explain how it differs from the trend component.
5. Use a fitted trend line and seasonal indices to forecast future values of a time series under both the additive and multiplicative models, and understand the limitations of such forecasts.
6. Apply simple exponential smoothing and understand its role in short-term forecasting when the series does not exhibit a strong seasonal pattern.

9.1 INTRODUCTION

In all the preceding units of this book, we have assumed that the observations we analyze are independent of one another. The marks of one student in a class, the income of one household in a city, and the response of one patient to a treatment have all been treated as unrelated to the others. In many real-world research situations, however, this assumption of independence is not tenable. When we collect data over time — for example, the monthly sales of a company in Islamabad over the past five years, the annual production of wheat in Punjab, or the daily closing price of a stock listed on the Pakistan Stock Exchange — the observations are almost always correlated with one another. Such data, collected at successive and equally spaced intervals of time, are called a time series.

The analysis of time series data requires a different set of techniques from those we have studied so far. The goal of time series analysis is to understand the underlying structure of such data — to decompose it into its component parts, to identify the patterns that repeat over time, and to use those patterns to make informed forecasts about the future.

The present unit introduces the basic concepts and methods of time series analysis and forecasting. We begin in Sections 9.2 and 9.3 by defining a time series and describing its four main components: the secular trend (T), the cyclical variation (C), the seasonal variation (S), and the irregular variation (I). Section 9.4 introduces the additive and multiplicative decomposition models.

Sections 9.5 to 9.9 present four methods for measuring the secular trend: the freehand method, the method of semi-averages, the method of moving averages, and the method of least squares. Sections 9.10 to 9.12 present two methods for measuring seasonal variations. Section 9.13 explains deseasonalization; Section 9.14 covers the cyclical component; Section 9.15 brings everything together for forecasting; and Section 9.16 introduces exponential smoothing methods.

9.2 THE MEANING AND USES OF TIME SERIES ANALYSIS

A time series is a set of observations of a variable recorded at successive and equally spaced intervals of time. Common examples in the social sciences include:

- The monthly sales revenue of a retail store in Lahore over the past five years.
- The quarterly gross domestic product (GDP) of Pakistan from 2000 to the present.
- The daily number of patients admitted to a hospital in Karachi during a particular year.
- The annual production of wheat in Punjab from 1990 to 2020.
- The weekly closing price of a stock listed on the Pakistan Stock Exchange.
- The monthly unemployment rate in Pakistan over the past decade.

Time series analysis has two main objectives. The first is to understand the past behavior of the data by decomposing a series into its component parts — trend, cycle, seasonal pattern, and irregular fluctuations. The second, often of more immediate practical interest, is to forecast the future. Forecasting is not about certainty; it is about providing a range of plausible values that decision-makers can use as a basis for planning.

A time series model is a purely empirical description of past behavior projected into the future. It does not explain why the variable behaves as it does; it simply extrapolates observed patterns. Time series forecasting is most reliable in the short term and in stable conditions; it is much less reliable for long-term forecasts or when underlying conditions are subject to sudden, unpredictable change.

9.3 THE FOUR COMPONENTS OF A TIME SERIES

A time series is generally believed to be the result of the combined action of four distinct types of forces: the secular trend (T), the cyclical variation (C), the seasonal variation (S), and the irregular variation (I). The observed value at any point in time is assumed to be a function of these four components.

9.3.1 Secular Trend (T)

The secular trend is the long-term direction in which the time series is moving over a period of typically ten years or more. The trend may be upward, downward, or stationary. It is the most important component for long-range planning and for understanding the overall direction of the data. Examples include the steadily rising population of Pakistan, the increasing literacy rate, and the sharp upward trend in mobile-phone subscribers during the 2000s.

9.3.2 Cyclical Variation (C)

The cyclical variation consists of medium-term oscillations around the trend line with a duration of more than one year. Unlike seasonal variation, the cyclical variation does not have a fixed period; cycles can vary in length and amplitude. The most familiar example is the business cycle — the pattern of expansion, contraction, recession, and recovery. A typical business cycle may last from two to ten years. A period when the series is above the trend is called a peak or expansion; a period below the trend is called a trough or contraction.

9.3.3 Seasonal Variation (S)

The seasonal variation is the regular, periodic pattern that repeats itself every year, caused by factors operating on a yearly cycle: the seasons of the year, the months, the weeks, and so on. Sales of air conditioners peak in summer; sales of heaters peak in winter; textbook sales peak before the semester. The key feature is that seasonal variation is regular and predictable — the same months show the same direction of movement relative to the annual average, year after year.

9.3.4 Irregular Variation (I)

The irregular variation (also called the residual or random component) is the part of the time series that cannot be explained by trend, cycle, or seasonal variation. It consists of short-term, unpredictable fluctuations caused by random events: strikes, unexpected policy changes, natural disasters, or simply the inherent randomness of real-world processes. The irregular component is, by definition, unpredictable; no forecast can be perfect, which is why we must always speak in terms of probabilities and confidence intervals.

9.4 TIME SERIES DECOMPOSITION MODELS: ADDITIVE AND MULTIPLICATIVE

The four components T, C, S, and I are not directly observed. Only the actual series Y is observed, assumed to result from combining these four components. The two most common models are the additive model and the multiplicative model.

9.4.1 The Additive Model

In the additive model, the four components add together to produce the observed value:

$$Y = T + C + S + I \quad (9.1)$$

The additive model is appropriate when the magnitude of the seasonal variation does not change as the level of the series changes — that is, when the seasonal pattern is constant in absolute terms. For example, if sales

are always Rs. 10,000 higher in December regardless of the overall level, the additive model is appropriate. It is simpler to work with because all components are measured in the same units as the original data.

9.4.2 The Multiplicative Model

In the multiplicative model, the four components multiply together:

$$Y = T \times C \times S \times I \quad (9.2)$$

The multiplicative model is appropriate when the magnitude of the seasonal variation is proportional to the level of the series — that is, the seasonal pattern is constant in relative (percentage) terms. For example, if sales are always 20% higher in December, the multiplicative model applies. This model is more realistic for most economic and business time series because variability tends to increase with the level. Taking logarithms of both sides converts equation (9.2) into an additive model in log-scale:

$$\log Y = \log T + \log C + \log S + \log I \qquad (9.3)$$

9.4.3 Which Model to Choose?

Inspect a plot of the time series. If seasonal fluctuations appear roughly constant in magnitude over time, the additive model is appropriate. If the fluctuations increase as the series trends upward (peaks grow higher and troughs grow deeper), the multiplicative model is appropriate. A useful diagnostic is to check whether the ratio of seasonal amplitude to average level is roughly constant; if so, the multiplicative model is indicated. For most economic and business data, the multiplicative model is the standard choice.

9.5 MEASURING THE SECULAR TREND: AN OVERVIEW

The first step in time series decomposition is to isolate and measure the secular trend T. Once measured, it can be removed from the original series to reveal the other components. The four principal methods are:

1. The freehand (graphical) method.
2. The method of semi-averages.
3. The method of moving averages.
4. The method of least squares.

The freehand and semi-average methods are relatively crude and subjective, suitable only for a quick rough description. Moving averages are widely used for smoothing. Least squares is the most precise and objective, providing a mathematical equation for forecasting. Each method is discussed in the following sections.

9.6 THE FREEHAND (GRAPHICAL) METHOD OF TREND MEASUREMENT

The freehand method is the simplest and most subjective approach. The time series is plotted on a graph with time on the horizontal axis and the variable on the vertical axis. A smooth curve is then drawn through the plotted points, following the general direction while ignoring short-term fluctuations.

9.6.1 The Procedure

1. Plot the time series on a graph, with time periods equally spaced on the horizontal axis and observed values on the vertical axis.
2. Examine the plotted points and decide on the general direction: increasing, decreasing, or stationary.
3. Draw a smooth curve (or straight line) through the points, following the overall movement while smoothing over short-term fluctuations. Aim to have roughly as many points above the curve as below it.

4. The resulting curve is the estimated trend.



Figure 9.1 : Freehand trend of annual sales for a Faisalabad textile firm (2015–2025)

9.6.2 Limitations of the Freehand Method

The freehand method is highly subjective: two analysts will almost certainly produce different trend lines on the same data. Because it is subjective, no measure of reliability can be attached to the estimated trend. The method is also unsuitable for forecasting, since there is no mathematical equation to extrapolate. For these reasons, the freehand method is rarely used in serious statistical work; it is mainly useful for a quick visual impression of the data.

Example 9.1 — Freehand Method

The following data show the annual sales (in million rupees) of a textile firm in Faisalabad from 2015 to 2025. Plot the data and describe the freehand trend.

Year	Sales	Year	Sales	Year	Sales
2015	45	2019	62	2023	78
2016	48	2020	65	2024	82
2017	52	2021	70	2025	86
2018	55	2022	74	—	—

Solution. The graph shows a steady upward movement from 45 in 2015 to 86 in 2025. A freehand trend line drawn through the points would be approximately straight with a positive slope (as shown in figure 9.1 above), passing through the middle of the scatter. Different analysts would draw slightly different lines. The method gives a quick visual impression of the trend but is not precise enough for formal forecasting.

9.7 THE METHOD OF SEMI-AVERAGES FOR TREND MEASUREMENT

The method of semi-averages is a simple, objective way to fit a straight-line trend. The series is divided into two equal halves, the arithmetic mean of each half is computed, and the two means are plotted at the mid-points of the respective half-periods. The line joining these two points is the estimated trend line.

9.7.1 The Procedure

1. Divide the time series into two equal halves. If the number of observations is odd, omit the middle observation.
2. Compute the arithmetic mean of the first half and the arithmetic mean of the second half.
3. Determine the mid-point of the time period for each half. For example, if the first half covers 2015–2019, the mid-point is the year 2017 (middle year).
4. Plot the two points (mid-point time, half-mean) and draw a straight line through them. This is the estimated trend line.
5. If a mathematical equation is desired, compute the slope and intercept from the two plotted points.

9.7.2 Merits and Demerits

The method of semi-averages is objective, easy to apply, and provides a straight-line trend with a mathematical equation suitable for forecasting. It is far less subjective than the freehand method.

The main limitation is that it assumes the trend is perfectly linear. If the trend is curved, a straight-line fit will be a poor description. The method also averages over cyclical and seasonal fluctuations within each half, which can bias the result if the two halves happen to cover different phases of the cycle. Despite these limitations, the method is useful for a quick approximation.

Example 9.2 — Method of Semi-Averages

Using the sales data of Example 9.1 (2015–2025), fit a straight-line trend by the method of semi-averages and forecast the sales for 2026.

Solution. The data cover 11 years. Since n is odd, omit the middle year (2020) and divide the remaining 10 years into two halves of 5 years each:

First half: 2015–2019 (Sales: 45, 48, 52, 55, 62)
Second half: 2021–2025 (Sales: 70, 74, 78, 82, 86)

Mean of first half: (45 + 48 + 52 + 55 + 62) / 5 = 262 / 5 = 52.4. Mid-point: 2017.

Mean of second half: (70 + 74 + 78 + 82 + 86) / 5 = 390 / 5 = 78.0. Mid-point: 2023.

Two points: (2017, 52.4) and (2023, 78.0).

$b = (78.0 - 52.4) / (2023 - 2017) = 25.6 / 6 = 4.267$ *million rupees per year*

$a = 52.4 - 4.267 \times 2017 = 52.4 - 8604.14 = -8551.74$

Trend equation: Trend = −8551.74 + 4.267 × year

Forecast for 2026:

$Trend(2026) = -8551.74 + 4.267 \times 2026 = -8551.74 + 8645.94 = 94.20$ *million rupees*

The forecast is Rs. 94.20 million. This is a linear extrapolation assuming the trend continues at Rs. 4.267 million per year. Caution is warranted when extrapolating beyond the observed data.

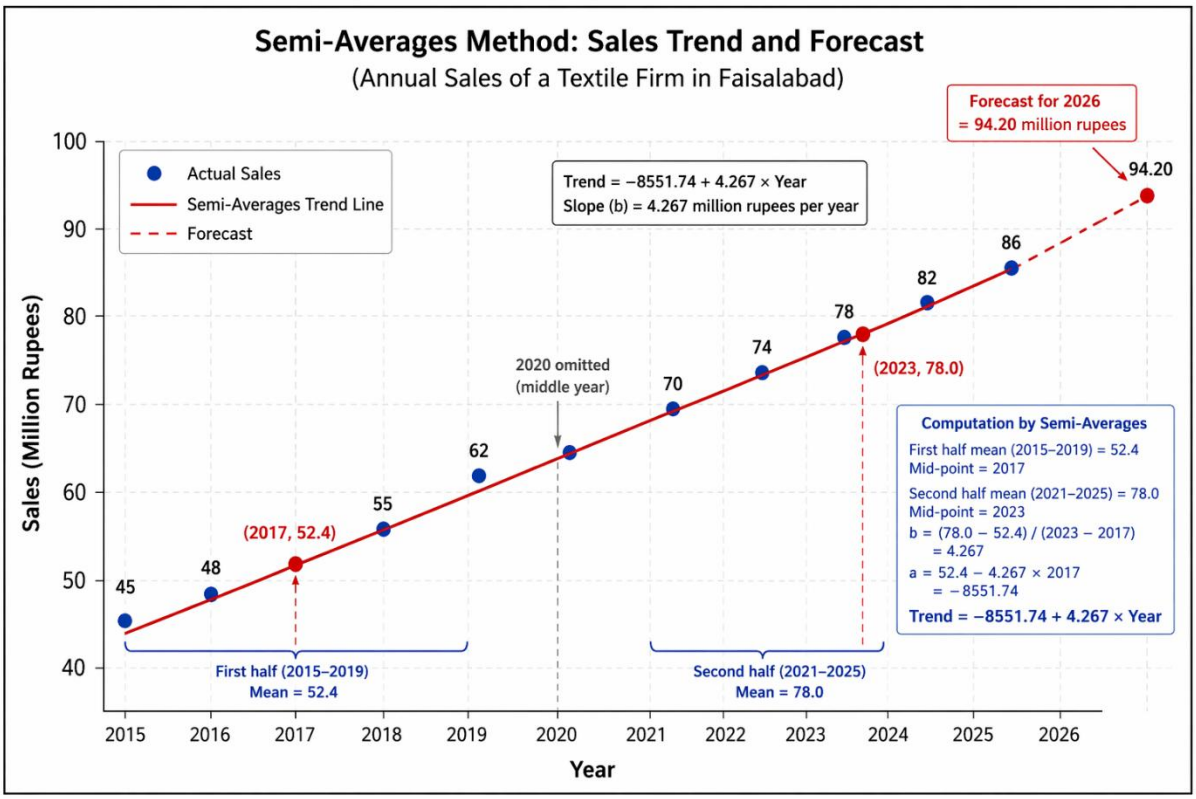


Figure 9.2: Semi-averages trend line and 2026 forecast of annual sales (2015–2025)

9.8 THE METHOD OF MOVING AVERAGES FOR TREND MEASUREMENT

The method of moving averages is one of the most widely used techniques for smoothing a time series and measuring the trend. A moving average of order n replaces each observation with the average of that observation and a certain number of its neighbors, producing a smoother series in which short-term fluctuations are reduced or eliminated.

9.8.1 The Procedure

For an odd-order moving average (n = 3, 5, 7, …), compute the average of the first n observations and place it at the middle of the period covered. Then drop the first observation, add the (n+1)-th, compute the new average, and place it at the middle of the next period. Continue until the series is covered.

When n is even (n = 4, 12), the moving averages fall between time periods and do not align directly with the original data. In this case, compute centred moving averages by averaging pairs of successive moving averages. For quarterly data (n = 4), first compute 4-quarter moving totals, then average them in pairs to obtain centred moving averages that align with the original quarters.

9.8.2 Choosing the Order of the Moving Average

The order n should be chosen to equal the length of the cycle to be eliminated. For monthly data, use a 12-month moving average to eliminate seasonal variation; for quarterly data, use a 4-quarter moving average. A lower-order moving average follows the data more closely but retains more short-term variation. A higher-order moving average is smoother but loses more observations at the ends of the series: for odd n, the first and last $(n - 1)/2$ observations cannot have moving averages computed; for even n, the first and last $n/2$ observations are lost.

Example 9.3 — Moving Average Method (4-Quarter Centred)

The following data show the quarterly production (thousand units) of a cement factory in Lahore from 2020 to 2022. Compute the 4-quarter centred moving averages to estimate the trend.

Year	Q I	Q II	Q III	Q IV
2020	120	130	125	140
2021	135	145	140	155
2022	150	160	155	170

Solution. Compute 4-quarter moving totals, then average consecutive totals (divide by 8) to obtain centred moving averages (CMA) aligned with each quarter:

Year	Qtr	Production	4-Qtr Total	4-Qtr MA	CMA (Trend)
2020	I	120	—	—	—
2020	II	130	—	—	—
2020	III	125	515	128.75	130.625
2020	IV	140	530	132.50	133.750
2021	I	135	545	136.25	138.125
2021	II	145	560	140.00	141.250
2021	III	140	575	143.75	145.625
2021	IV	155	590	147.50	148.750
2022	I	150	605	151.25	153.125
2022	II	160	620	155.00	156.250
2022	III	155	635	158.75	—
2022	IV	170	—	—	—

For example: 4-quarter total for Q III 2020 = 120 + 130 + 125 + 140 = 515; 4-quarter MA = 515/4 = 128.75 (placed between Q II and Q III 2020). CMA for Q III 2020 = (128.75 + 132.50)/2 = 130.625. The centred moving averages are the estimated trend values and clearly show the upward trend in production.

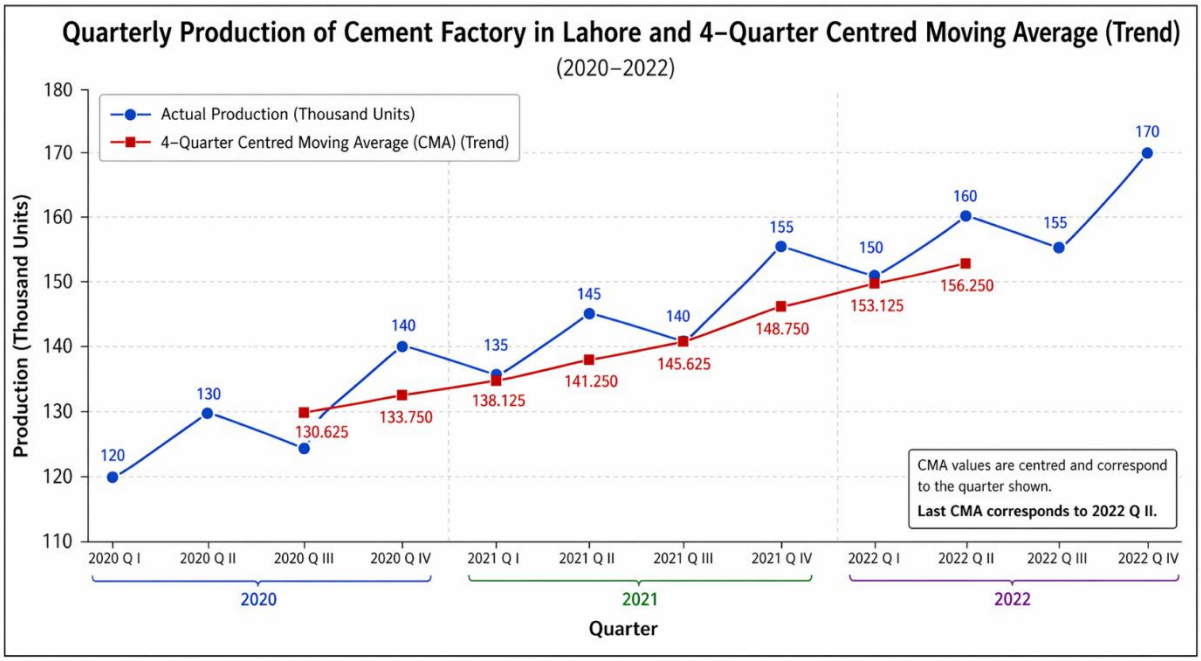


Figure 9.3: Quarterly cement production (2020–2022) with 4-quarter centred moving averages

9.9 THE METHOD OF LEAST SQUARES FOR TREND MEASUREMENT

The method of least squares is the most precise and objective method for fitting a trend line. It minimizes the sum of squared vertical deviations of the observed values from the fitted line. The fitted line can then be used to forecast future values.

9.9.1 Fitting a Linear Trend

Let Y_t denote the value of the time series at time t (t = 1, 2, …, n). The linear trend model is:

$$Y_t = a + b\,t + \varepsilon_t \qquad (9.4)$$

where a is the intercept, b is the slope (average increase per unit time), and ε_t is the random error. The least squares estimates of a and b are:

$$b = [n\,\Sigma(t\,Y_t) - (\Sigma t)(\Sigma Y_t)] \,/\, [n\,\Sigma t^2 - (\Sigma t)^2] \qquad (9.5)$$

$$a = \bar{Y} - b\,\bar{t} \qquad (9.6)$$

where $\bar{t}$ and $\bar{Y}$ are the sample means of t and Y.

A convenient simplification is to code time so that $\Sigma t = 0$. For odd n, set t = …, −2, −1, 0, 1, 2, …. When $\Sigma t = 0$, the formulas reduce to:

$$b = \Sigma(t\,Y) \,/\, \Sigma(t^2)\,, \quad a = \bar{Y} \qquad (9.7)$$

This simplification greatly reduces the arithmetic involved and is the preferred approach for hand calculation.

9.9.2 The Standard Error of Estimate

The standard error of estimate measures the typical vertical distance between the observed values and the fitted trend line:

$$s_e = \sqrt{[\,\Sigma(Y_t - \hat{Y}_t)^2 \,/\, (n-2)\,]} \qquad (9.8)$$

A smaller s_e indicates a better fit. It is also used to construct prediction intervals for forecasts.

Example 9.4 — Least Squares Linear Trend

Fit a linear trend by the method of least squares to the annual sales data of Example 9.1 (2015–2025) and forecast the sales for 2026.

Solution. Let t = 1, 2, …, 11 correspond to 2015, 2016, …, 2025.

Year	t	Sales Y	tY	t^2	$\hat{Y}$ (Fitted)
2015	1	45	45	1	44.00
2016	2	48	96	4	48.27
2017	3	52	156	9	52.55
2018	4	55	220	16	56.82
2019	5	62	310	25	61.09
2020	6	65	390	36	65.36
2021	7	70	490	49	69.64
2022	8	74	592	64	73.91
2023	9	78	702	81	78.18
2024	10	82	820	100	82.45
2025	11	86	946	121	86.73
Total	66	717	4767	506	—

$\bar{t} = 66/11 = 6, \quad \bar{Y} = 717/11 = 65.1818$

$b = [11(4767) - (66)(717)] / [11(506) - (66)^2] = [52437 - 47322] / [5566 - 4356] = 5115 / 1210 = 4.2273$

$a = 65.1818 - 4.2273 \times 6 = 65.1818 - 25.3638 = 39.8180$

Fitted trend equation:

$\hat{Y}_t = 39.82 + 4.23\,t \quad (t = 1 \text{ for } 2015)$

Forecast for 2026 (t = 12):

$\hat{Y}_{12} = 39.82 + 4.2273 \times 12 = 39.82 + 50.73 = 90.55$ *million rupees*

The trend increases at approximately Rs. 4.23 million per year. The forecast of Rs. 90.55 million for 2026 is a linear extrapolation and should be interpreted cautiously.

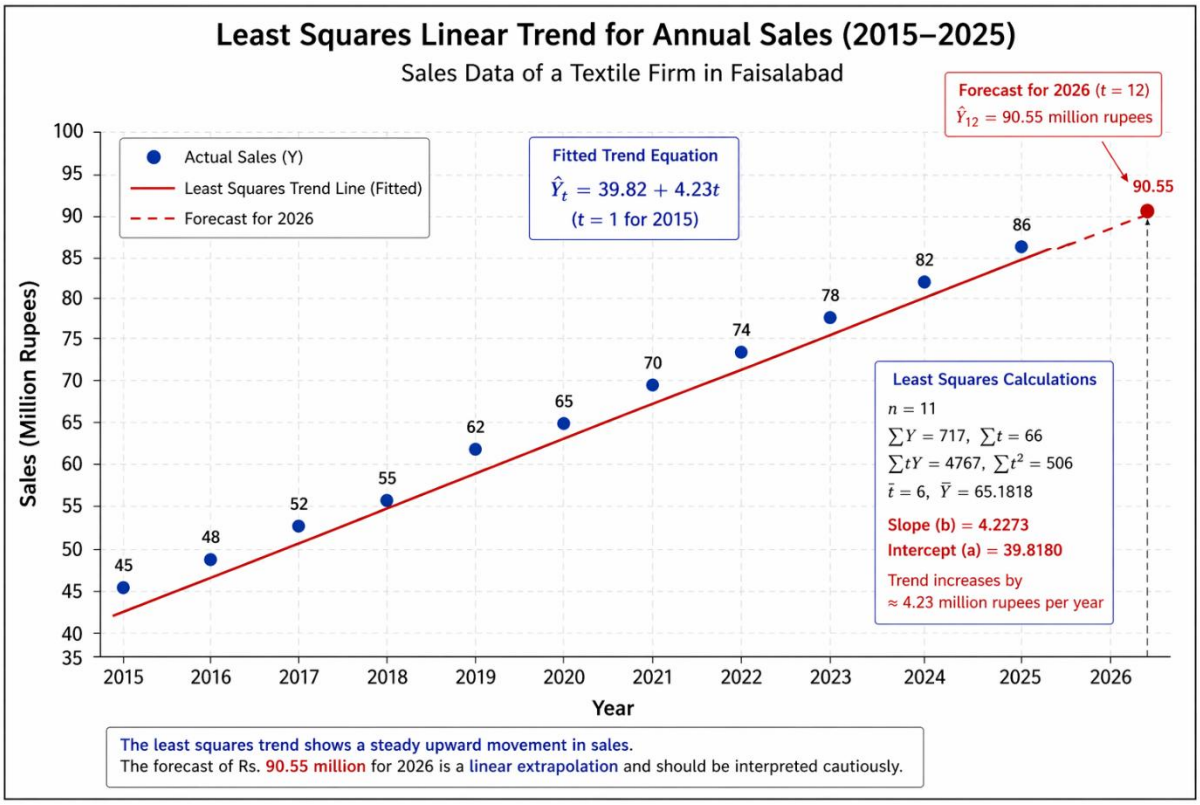


Figure 9.3: Least squares linear trend fitted to annual sales (2015–2025) with 2026 forecast

9.10 MEASURING SEASONAL VARIATIONS: AN OVERVIEW

After measuring and removing the trend from a time series, the next step is to measure the seasonal variation. Seasonal indices are numbers that describe the typical deviation of each season (quarter or month) from the annual average, after the trend has been removed. A seasonal index of 120 for a particular quarter means that, on average, the value in that quarter is 20% above the annual average (after accounting for trend). An index of 85 means the value is 15% below average.

Two principal methods are available for measuring seasonal variations: the ratio-to-trend method (Section 9.11) and the ratio-to-moving-average method (Section 9.12). The ratio-to-moving-average method is more widely used and is the preferred method in official statistics.

9.11 THE RATIO-TO-TREND METHOD FOR MEASURING SEASONAL VARIATIONS

The ratio-to-trend method requires that a trend line has already been fitted to the time series. The method then isolates the seasonal and irregular components by computing, for each period, the ratio of the actual value to the trend value.

9.11.1 Procedure

1. Compute the trend values $\hat{Y}_t$ for each time period using the fitted trend line.
2. For each period, compute the ratio of the actual value to the trend value (multiplicative model):

   $$R = Y / T \qquad (9.8)$$

3. These ratios capture the seasonal and irregular components (S × I).

4. For each season (quarter or month), compute the average of the ratios for that season across all years. Averaging over years largely cancels the irregular component, leaving the seasonal index for that season.
5. Adjust the seasonal indices so that their average equals 100 (sum = 400 for quarterly data; sum = 1200 for monthly data). Multiply each index by the adjustment factor (400 / sum of unadjusted indices) for quarterly data.

9.11.2 Interpreting Seasonal Indices

A seasonal index of 120 means the value in that season is, on average, 20% above the annual average after removing trend. An index of 85 means 15% below average. Seasonal indices are always expressed as percentages with 100 representing the average level.

Example 9.5 — Ratio-to-Trend Method

The quarterly sales (million rupees) of a department store in Islamabad for 2020–2022 are shown below. A linear trend has been fitted by least squares, giving the trend values shown. Compute the seasonal indices by the ratio-to-trend method.

Year	Quarter	Actual Y	Trend T	Ratio R = Y/T
2020	I	45	48.0	0.9375
2020	II	52	50.5	1.0297
2020	III	48	53.0	0.9057
2020	IV	55	55.5	0.9910
2021	I	50	58.0	0.8621
2021	II	58	60.5	0.9587
2021	III	55	63.0	0.8730
2021	IV	62	65.5	0.9466
2022	I	58	68.0	0.8529
2022	II	65	70.5	0.9219
2022	III	62	73.0	0.8493
2022	IV	70	75.5	0.9272

Solution. Average the ratios by quarter:

Q I: (0.9375 + 0.8621 + 0.8529) / 3 = 0.8842 → 88.42%

Q II: (1.0297 + 0.9587 + 0.9219) / 3 = 0.9701 → 97.01%

Q III: (0.9057 + 0.8730 + 0.8493) / 3 = 0.8760 → 87.60%

Q IV: (0.9910 + 0.9466 + 0.9272) / 3 = 0.9549 → 95.49%

Sum of averages = 88.42 + 97.01 + 87.60 + 95.49 = 368.52. Adjustment factor = 400 / 368.52 = 1.0854. Adjusted seasonal indices:

Q I: 88.42 × 1.0854 ≈ 96.0%

Q II: 97.01 × 1.0854 ≈ 105.3%

Q III: 87.60 × 1.0854 ≈ 95.1%

Q IV: 95.49 × 1.0854 ≈ 103.6%

Sum = 400.0 (after rounding adjustments). The second quarter is the strongest (105.3% of the annual average after removing trend) and the third quarter is the weakest (95.1%).

9.12 THE RATIO-TO-MOVING-AVERAGE METHOD FOR MEASURING SEASONAL VARIATIONS

The ratio-to-moving-average method is the most widely used method for measuring seasonal variations. It does not require a pre-fitted trend line; instead, it uses centred moving averages to estimate the trend-cycle component (T × C), which is then removed from the data to leave the seasonal-irregular component (S × I).

9.12.1 Procedure

1. Compute a centred moving average (CMA) of the time series using a period equal to the number of seasons per year (4 for quarterly data, 12 for monthly data). The CMA estimates the trend-cycle component T × C.
2. For each period with a CMA available, compute the ratio: (Y / CMA) = S × I.
3. For each season, average the ratios for that season across all years. Averaging over years largely cancels the irregular component, leaving the seasonal index for that season.
4. Adjust the seasonal indices so that their sum equals 400 (quarterly) or 1200 (monthly), by multiplying each by the appropriate correction factor.

9.12.2 Advantages of the Ratio-to-Moving-Average Method

This method is superior to the ratio-to-trend method because it does not require a prior estimate of the trend. The moving average automatically smooths out seasonal and irregular components, leaving the trend-cycle as the base for computing seasonal ratios. It is the basis for the seasonal adjustment procedures used by many government statistical agencies worldwide.

Example 9.6 — Ratio-to-Moving-Average Method

Using the quarterly production data of Example 9.3, compute the seasonal indices by the ratio-to-moving-average method.

Solution. The centred moving averages (CMA) were computed in Example 9.3. Compute the ratios Y/CMA for each available quarter:

Year	Qtr	Y	CMA	Ratio (S×I)	Quarter Avg → Adj. SI
2020	I	120	—	—	Q I: (0.977+0.980)/2 = 0.979 → 97.9%
2020	II	130	—	—	Q II: (1.026+1.024)/2 = 1.025 → 102.5%
2020	III	125	130.625	0.957	Q III: (0.957+0.961)/2 = 0.959 → 95.9%
2020	IV	140	133.750	1.047	Q IV: (1.047+1.042)/2 = 1.045 → 104.3%
2021	I	135	138.125	0.977	—
2021	II	145	141.250	1.026	—
2021	III	140	145.625	0.961	—
2021	IV	155	148.750	1.042	—
2022	I	150	153.125	0.980	—
2022	II	160	156.250	1.024	—
2022	III	155	—	—	—

2022	IV	170	—	—	—

Quarter averages: Q I = 97.85%, Q II = 102.50%, Q III = 95.90%, Q IV = 104.45%. Sum = 400.70.

Adjustment factor = 400 / 400.70 ≈ 0.9983. Adjusted seasonal indices (rounded): Q I ≈ 97.7%, Q II ≈ 102.3%, Q III ≈ 95.7%, Q IV ≈ 104.3%.

Interpretation: Production in Q IV is about 4.3% above the annual average (after removing trend and cycle), while Q III is about 4.3% below. These indices can be used to deseasonalize the data and to forecast future production.

9.13 DESEASONALIZING A TIME SERIES

Deseasonalization removes the seasonal component from a time series so that the underlying trend and cyclical movements can be seen more clearly. Once the seasonal indices have been computed, the deseasonalized values are obtained by dividing the original values by the seasonal indices (in the multiplicative model) or by subtracting the seasonal component (in the additive model).

9.13.1 Procedure for Deseasonalization

In the multiplicative model, the deseasonalized value for each period is:

$$\textit{Deseasonalized value} = \textit{Actual value} / \textit{Seasonal index (decimal form)} \quad (9.9)$$

For example, if the actual value is 140 and the seasonal index for that quarter is 1.04 (104%), the deseasonalized value is 140 / 1.04 = 134.62. This represents what the value would have been in the absence of seasonal variation.

9.13.2 Why Deseasonalize?

Deseasonalized data are much easier to interpret than the original data because the regular seasonal fluctuations have been removed. The underlying trend and cyclical movements are visible without the distraction of the seasonal pattern. Deseasonalized data are also used as inputs to other statistical procedures such as regression analysis, where seasonal effects would otherwise distort the results.

Example 9.7 — Deseasonalizing a Time Series

Using the seasonal indices from Example 9.5 (Q I = 96.0%, Q II = 105.3%, Q III = 95.1%, Q IV = 103.6%), deseasonalize the actual quarterly sales for 2022. The actual sales for 2022 were: Q I = 58, Q II = 65, Q III = 62, Q IV = 70 million rupees.

Solution. Convert seasonal indices to decimal form: Q I = 0.960, Q II = 1.053, Q III = 0.951, Q IV = 1.036.

Deseasonalized Q I = 58 / 0.960 = 60.42 million rupees

Deseasonalized Q II = 65 / 1.053 = 61.73 million rupees

Deseasonalized Q III = 62 / 0.951 = 65.19 million rupees

Deseasonalized Q IV = 70 / 1.036 = 67.57 million rupees

The deseasonalized series (60.42, 61.73, 65.19, 67.57) shows a steady upward movement, reflecting the underlying trend and cyclical factors. The seasonal pattern has been removed, making the trend much easier to see.

9.14 MEASURING CYCLICAL VARIATION

The cyclical component represents medium-term oscillations around the trend with durations of more than one year. The most common method for measuring it is the residual method.

9.14.1 The Residual Method

Once the trend and seasonal components have been removed, the remaining variation is the combined effect of the cyclical and irregular components (C × I). To isolate the cyclical component, we smooth the (C × I) series by applying a moving average. The resulting smoothed series is the estimate of the cyclical component C. The steps are:

1. Deseasonalize the original series (if seasonal variation is present).
2. Fit a trend line to the deseasonalized series and compute the trend values.
3. Compute the ratio (deseasonalized value / trend value) × 100. This ratio represents C × I.
4. Smooth the ratios using a moving average of an appropriate length (typically 3 or 5 years for annual data, or a period corresponding to the expected length of the business cycle). The resulting smoothed series is the cyclical component C.

9.14.2 Interpreting the Cyclical Component

The cyclical component is usually expressed as a percentage of the trend. A value of 105 means the series is 5% above the trend at that point; a value of 95 means it is 5% below. Peaks in the cyclical component indicate periods of expansion; troughs indicate periods of contraction. The cyclical component is of great interest to economists and business planners because it signals the phases of the business cycle.

9.15 FORECASTING WITH TIME SERIES MODELS

The ultimate goal of much time series analysis is forecasting — predicting future values on the basis of past behavior. Once the components have been measured, they are combined to produce forecasts. The method depends on whether the additive or multiplicative model is used.

9.15.1 Forecasting with the Multiplicative Model

In the multiplicative model, the forecast for a future period is:

$$\hat{Y} = \hat{T} \times \hat{C} \times \hat{S} \qquad (9.10)$$

where $\hat{T}$ is the forecasted trend value (by extrapolating the trend line), $\hat{C}$ is the forecasted cyclical component (often assumed to be 1.0 for short-term forecasts, since the cycle is difficult to predict), and $\hat{S}$ is the seasonal index for that period (in decimal form). For short-term forecasts, the formula reduces to:

$$\hat{Y} = \hat{T} \times \hat{S} \qquad (9.11)$$

9.15.2 Forecasting with the Additive Model

In the additive model, the forecast is:

$$\hat{Y} = \hat{T} + \hat{C} + \hat{S} \qquad (9.12)$$

For short-term forecasts, the cyclical component is often assumed to be zero, giving:

$$\hat{Y} = \hat{T} + \hat{S} \qquad (9.13)$$

9.15.3 Limitations of Time Series Forecasting

Time series forecasts are based solely on the extrapolation of past patterns. They do not account for changes in the underlying conditions that may cause the future to differ from the past. A forecast is reliable only to the extent that the historical pattern continues. If there is a structural break — a sudden, permanent change in the behavior of the series — the forecast will be wrong. Time series forecasts should always be interpreted with caution and updated frequently as new data become available.

Example 9.8 — Forecasting with the Multiplicative Model

Using the least squares trend line from Example 9.4 ($\hat{Y}_t = 39.82 + 4.2273\ t$) and the seasonal indices from Example 9.5 (Q I = 96.0%, Q II = 105.3%, Q III = 95.1%, Q IV = 103.6%), forecast the quarterly sales for the four quarters of 2023.

Solution. The annual trend line uses t = 1 for 2015, …, t = 11 for 2025. For quarterly forecasting, convert the annual trend to a quarterly basis. The annual increment is 4.2273 million per year; the quarterly increment is 4.2273/4 = 1.0568 million per quarter. For illustration, assume a quarterly trend has been fitted that gives the following trend values for 2023: Q I = 90.0, Q II = 91.0, Q III = 92.0, Q IV = 93.0 million rupees.

Using equation (9.11) with seasonal indices in decimal form:

Q I: 90.0 × 0.960 = 86.4 million rupees

Q II: 91.0 × 1.053 = 95.8 million rupees

Q III: 92.0 × 0.951 = 87.5 million rupees

Q IV: 93.0 × 1.036 = 96.3 million rupees

These forecasts incorporate both the upward trend and the seasonal pattern. Q II and Q IV are forecast higher due to their larger seasonal indices; Q I and Q III are forecast lower.

9.16 EXPONENTIAL SMOOTHING METHODS FOR FORECASTING

Exponential smoothing is a family of forecasting methods that assign exponentially decreasing weights to past observations, with the most recent observation receiving the highest weight. These methods are particularly useful for short-term forecasting when the series does not exhibit a strong or measurable seasonal pattern.

9.16.1 Simple Exponential Smoothing (SES)

Simple exponential smoothing is used for time series that have no systematic trend or seasonal pattern. The forecast for the next period is:

$$F_{t+1} = \alpha\, Y_t + (1 - \alpha)\, F_t \qquad (9.14)$$

where F_t is the forecast for period t, Y_t is the actual value in period t, and α (alpha) is the smoothing constant, $0 < \alpha < 1$. A high α gives more weight to recent observations, making the forecast more responsive to changes. A low α gives more weight to past observations, making the forecast more stable. The optimal value of α is chosen to minimize the mean absolute deviation (MAD) or mean squared error (MSE). In practice, α is often set between 0.05 and 0.30, with α = 0.1 or 0.2 being common choices.

9.16.2 Holt's Linear Exponential Smoothing (Double Exponential Smoothing)

Holt's method extends simple exponential smoothing to handle time series with a linear trend. It uses two smoothing constants: α for the level and β for the trend. Both the level and the slope are updated at each period, allowing the forecast to track a changing trend.

9.16.3 Holt-Winters Exponential Smoothing (Triple Exponential Smoothing)

The Holt-Winters method further extends exponential smoothing to handle time series that have both a trend and a seasonal pattern. Three smoothing constants are used: α for the level, β for the trend, and γ (gamma) for the seasonal component. The Holt-Winters method is one of the most widely used forecasting methods for seasonal time series in business and economics. Full details of these methods are beyond the scope of this course; the interested student is referred to the suggested readings.

Example 9.9 — Simple Exponential Smoothing

The monthly sales (thousand rupees) of a small electronics shop in Rawalpindi for the first six months of a year are given below. Using simple exponential smoothing with $\alpha = 0.2$, forecast the sales for July. The initial forecast for January is $F_1 = 50$.

Month	Actual Sales Y_t	Forecast F_t
January	52	50 (given)
February	55	$F_2 = 0.2(52) + 0.8(50) = 50.40$
March	53	$F_3 = 0.2(55) + 0.8(50.40) = 51.32$
April	58	$F_4 = 0.2(53) + 0.8(51.32) = 51.66$
May	60	$F_5 = 0.2(58) + 0.8(51.66) = 52.92$
June	62	$F_6 = 0.2(60) + 0.8(52.92) = 54.34$
July	—	$F_7 = 0.2(62) + 0.8(54.34) = 55.87$

The exponential smoothing forecast for July is approximately Rs. 55,870. Notice how the forecast gradually adjusts upward as actual sales increase, but it lags behind the actual values because of the smoothing effect. A higher value of α would make the forecast more responsive to recent increases.

9.17 CHOOSING THE RIGHT FORECASTING METHOD: A DECISION MAP

With several methods available, the following decision map summarizes how to choose:

1. What is your goal? If you simply want to describe past behavior, any trend-measurement method may be adequate, but least squares is preferred for its objectivity and precision.
2. Is there a seasonal pattern? If the data are annual, focus on trend and cyclical components only. If the data are quarterly or monthly, measure the seasonal component before examining the underlying trend.
3. How much data do you have? Moving-average and ratio-to-moving-average methods require enough data to compute moving averages without losing too many observations at the ends. For short time series (fewer than about 5 years of quarterly data), the ratio-to-trend method may be more appropriate.
4. Do you need a mathematical equation for forecasting? If yes, use least squares for the trend and combine with seasonal indices to produce forecasts.
5. Is the series relatively stable with no strong trend or seasonality? If yes, simple exponential smoothing may be an appropriate and easy-to-implement forecasting method.
6. Do you need to update forecasts frequently as new data arrive? Exponential smoothing methods are ideal: they can be updated with a single calculation each period, without re-estimating the entire model.

9.18 COMMON PITFALLS IN TIME SERIES ANALYSIS

9.18.1 Extrapolating Too Far Beyond the Data

The most common error in time series forecasting is to extrapolate the trend too far into the future. A linear trend fitted to 10 years of data may give a reasonable forecast one or two years ahead, but it becomes increasingly unreliable the further out we go. The pattern may change, growth may slow, or a turning point

may occur. Every forecast should be accompanied by a statement of its limitations, and long-term forecasts based on short-term data should be avoided.

9.18.2 Ignoring the Seasonal Pattern

When quarterly or monthly data are analyzed, the seasonal pattern must be taken into account. Comparing the sales of a retail store in December with its sales in July without adjusting for seasonality is meaningless, because the two months are not directly comparable. Deseasonalizing the data before making such comparisons is essential.

9.18.3 Using an Inappropriate Decomposition Model

Choosing between the additive and multiplicative models without inspecting the data can lead to incorrect seasonal indices. If the seasonal pattern is proportional to the level of the series, the multiplicative model is appropriate; if constant in absolute terms, the additive model is appropriate. A quick plot of the data will usually indicate which model is more suitable.

9.18.4 Confusing the Cyclical Component with the Trend

The cyclical component represents medium-term oscillations around the trend, while the trend represents the long-term direction. A downturn in the cyclical component does not mean the trend has reversed; it may simply be a temporary contraction within a long-term upward trend. These two must always be interpreted separately.

9.18.5 Overfitting the Model

Adding more parameters (e.g., fitting a quadratic or cubic trend instead of a linear trend) will always improve the fit to historical data, but it may lead to overfitting — capturing the random noise in the past rather than the underlying pattern. An overfitted model usually forecasts poorly. The principle of parsimony (using the simplest model that adequately describes the data) is as important in time series analysis as in any other area of statistics.

9.18.6 Forgetting that Correlation is Not Causation

Two time series may be highly correlated simply because both are trending upward over time, not because one causes the other (spurious correlation). For example, the number of drownings and the consumption of ice cream are both positively correlated with temperature, not with each other. The student should always be cautious about inferring causality from time series data.

9.19 SUMMARY

This unit has introduced the basic concepts and methods of time series analysis and forecasting. The main points are:

- A time series is a set of observations recorded at successive, equally spaced intervals of time. The analysis aims to understand past behavior and to forecast future values.
- A time series is decomposed into four components: the secular trend T (long-term direction), the cyclical variation C (medium-term oscillations), the seasonal variation S (regular yearly pattern), and the irregular variation I (random fluctuations). The additive model is $Y = T + C + S + I$; the multiplicative model is $Y = T \times C \times S \times I$.

- The secular trend is measured by the freehand method (subjective, graphical), the method of semi-averages (objective straight line), the method of moving averages (smoothing), or the method of least squares (most precise, provides a mathematical equation for forecasting).
- Seasonal variations are measured by the ratio-to-trend method or the ratio-to-moving-average method. The ratio-to-moving-average method is preferred because it does not require a pre-fitted trend line. Seasonal indices are expressed as percentages, with 100 representing the average level. For quarterly data, adjusted indices sum to 400; for monthly data, to 1200.
- Deseasonalization removes the seasonal component, allowing the underlying trend and cyclical movements to be seen more clearly. In the multiplicative model, deseasonalized values = actual values / seasonal indices (decimal form).
- The cyclical component is measured by the residual method: after removing trend and seasonal components, the remaining C × I series is smoothed (by moving average) to isolate C.
- Forecasting with the multiplicative model uses $\hat{Y} = \hat{T} \times \hat{S}$ for short-term forecasts (assuming $\hat{C} \approx 1$). For the additive model, $\hat{Y} = \hat{T} + \hat{S}$.
- Exponential smoothing methods — simple exponential smoothing (no trend or season), Holt's method (linear trend), and Holt-Winters method (trend and seasonality) — are useful for short-term forecasting and for updating forecasts efficiently as new data arrive.
- Common pitfalls include: extrapolating too far beyond the data, ignoring seasonality, using an inappropriate decomposition model, confusing cyclical with trend movements, overfitting, and confusing correlation with causation.

9.20 SELF-ASSESSMENT EXERCISE

1. Define a time series. List the four main components of a time series and give a brief description of each.
2. Distinguish between the additive and multiplicative decomposition models for a time series. Under what circumstances is each model appropriate?
3. Describe the method of moving averages for measuring the secular trend of a time series. How do you choose the order of the moving average?
4. Explain the ratio-to-moving-average method for measuring seasonal variations. Why is this method preferred over the ratio-to-trend method?
5. What is deseasonalization, and why is it useful? Show how to deseasonalize a time series under the multiplicative model.
6. Discuss in detail the four methods for measuring the secular trend of a time series: the freehand method, the method of semi-averages, the method of moving averages, and the method of least squares. For each method, describe the procedure, list the advantages and disadvantages, and give a practical situation in which the method would be most appropriate. Compare the methods in terms of objectivity, precision, and suitability for forecasting.
7. Describe the ratio-to-moving-average method for measuring seasonal variations in a time series. Provide a step-by-step procedure, including how to compute the centred moving averages, how to calculate the seasonal indices, and how to adjust the indices so that their sum is 400 for quarterly data. Illustrate with a complete numerical example of your own choosing.

8. A time series of quarterly data has seasonal indices of 80, 110, 95, and 115 for Q I, Q II, Q III, and Q IV respectively. The actual value in Q I of a particular year is 120. (a) What is the deseasonalized value for this quarter? (b) Interpret the seasonal index of 80 for Q I. (c) If the actual value in Q II is 150, what is the deseasonalized value? [Answers: (a) 150; (b) Q I is 20% below the annual average; (c) 136.36]

9. A time series has a linear trend $\hat{Y} = 50 + 2t$, where $t = 1$ for Q I 2020. The seasonal indices for the four quarters are 0.90, 1.10, 0.95, and 1.05. Forecast the value for Q IV 2022 ($t = 11$). [Answer: $\hat{Y} = 50 + 2(11) = 72$; forecast $= 72 \times 1.05 = 75.6$]

10. A monthly time series has a 12-month moving average of 150 for February 2023. The actual value for February 2023 is 180. (a) Compute the ratio of actual to moving average. (b) What does this ratio tell you about February? (c) What further step is needed to obtain the pure seasonal index for February? [Answers: (a) 1.20; (b) February is 20% above the trend-cycle level; (c) average the ratios for February across all available years]

11. A researcher fits a linear trend to a time series and obtains $R^2 = 0.85$. (a) What does this R^2 tell you about the fit? (b) If a quadratic trend gives $R^2 = 0.92$, should the researcher automatically prefer the quadratic model? (c) What principle should guide the choice of model complexity?

12. A time series from 2010 to 2025 shows a clear upward trend. Forecast for 2026 is made with a linear trend line, and also with a quadratic trend line that curves upward more sharply. (a) Which forecast is likely more accurate? (b) What additional information is needed to decide? (c) What is the danger of using a quadratic trend for forecasting beyond the observed data?

13. The annual production of sugar (thousand tonnes) in a Punjab factory from 2016 to 2025 is: 80, 85, 90, 95, 100, 105, 110, 115, 120, 125. Fit a linear trend by the method of least squares and forecast the production for 2026. [$\hat{Y} = 75.0 + 5.0t$; $t = 1$ for 2016; forecast for 2026 ($t = 11$) = 130 thousand tonnes]

14. Using the data from Problem 13, fit a linear trend by the method of semi-averages and compare the forecast for 2026 with the least squares forecast. [First half mean = 90 (mid-point 2018); second half mean = 115 (mid-point 2023); $b = 25/5 = 5.0$; forecast for 2026 = 130 — same as least squares]

15. The quarterly sales (million rupees) of a retail chain in Karachi for 2021–2023 are: 2021: 25, 28, 26, 30; 2022: 28, 32, 30, 35; 2023: 32, 36, 34, 38. Compute the 4-quarter centred moving averages. [Selected CMA values — Q III 2021: 27.25; Q IV 2021: 28.50; Q I 2022: 29.50; Q II 2022: 31.25; Q III 2022: 33.00; Q IV 2022: 34.50; Q I 2023: 35.50; Q II 2023: 36.50]

16. Using the quarterly data from Problem 15, compute the seasonal indices by the ratio-to-moving-average method. [Adjusted seasonal indices: Q I ≈ 97.4, Q II ≈ 101.3, Q III ≈ 96.9, Q IV ≈ 104.4]

17. Using the seasonal indices from Problem 16, deseasonalize the actual sales for the four quarters of 2023. [Q I: 32/0.974 = 32.85; Q II: 36/1.013 = 35.54; Q III: 34/0.969 = 35.09; Q IV: 38/1.044 = 36.40]

18. The monthly sales (thousand rupees) of a Sialkot business for the first six months of a year are: Jan 45, Feb 48, Mar 52, Apr 55, May 58, Jun 60. Using simple exponential smoothing with $\alpha = 0.3$ and initial forecast $F_1 = 46$, forecast July. [$F_2 = 45.7$; $F_3 = 47.6$; $F_4 = 48.9$; $F_5 = 50.8$; $F_6 = 53.0$; $F_7 \approx 55.1$ thousand rupees]

19. The annual GDP (billion rupees) of Pakistan from 2015 to 2025 (hypothetical) is: 2500, 2600, 2720, 2850, 3000, 3150, 3300, 3480, 3650, 3850, 4050. Fit a linear trend by least squares and compute the standard error of estimate. [$b \approx 155.0$; $a \approx 2272.7$; $\hat{Y} = 2272.7 + 155.0t$; $s_e \approx 45.2$ billion rupees]

20. Using the least squares trend from Problem 19, forecast GDP for 2026 and construct an approximate 95% prediction interval. [t = 12; $\hat{Y} = 2272.7 + 155.0(12) = 4132.7$; $t_{0.025,9} \approx 2.262$; PI $\approx 4132.7 \pm 2.262 \times 45.2 \approx (4030.5, 4234.9)$ billion rupees]

21. A quarterly time series has adjusted seasonal indices Q I = 90, Q II = 110, Q III = 95, Q IV = 105. The deseasonalized values for the four quarters of 2024 are: Q I = 100, Q II = 105, Q III = 110, Q IV = 115. Reconstruct the original actual values. [Q I = 90; Q II = 115.5; Q III = 104.5; Q IV = 120.75]

22. The annual sales of a company (million rupees) from 2018 to 2025 are: 120, 125, 130, 140, 145, 150, 160, 170. (a) Fit a linear trend by least squares. (b) Forecast for 2026 and 2027. (c) Comment on the reliability of the 2027 forecast. [$\hat{Y} = 112.5 + 6.9t$; 2026 (t = 9): 174.6; 2027 (t = 10): 181.5; forecast for 2027 is based on extrapolation two years beyond the data — moderate reliability]

23. A monthly time series has a 12-month moving average of 200 for January 2025. The actual value is 240. The average of all January ratios is 1.15. (a) What is the seasonal index for January? (b) Estimate the irregular component for January 2025. (c) Deseasonalize the January 2025 value. [(a) SI = 115%; (b) I = (240/200)/1.15 = 1.20/1.15 ≈ 1.0435; (c) deseasonalized = 240/1.15 ≈ 208.7]

24. The quarterly electricity consumption (million kWh) in a Punjab city for 2022–2024 is: 2022: 210, 230, 245, 220; 2023: 215, 235, 250, 225; 2024: 220, 240, 255, 230. Compute the seasonal indices by the ratio-to-trend method after fitting a linear trend to the quarterly data. [Approximate adjusted seasonal indices: Q I ≈ 99.5%, Q II ≈ 104.5%, Q III ≈ 105.5%, Q IV ≈ 90.5%]

25. Using the seasonal indices from Problem 24, forecast the electricity consumption for the four quarters of 2025, assuming the linear trend continues. [Extend the fitted quarterly trend line to 2025; multiply each trend value by the corresponding seasonal index in decimal form]

REFERENCES AND SUGGESTED READINGS

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Hoboken, NJ: John Wiley & Sons.

Chaudhry, S. M., & Kamal, S. (2010). *Introduction to statistical theory, Part II* (8th ed.). Lahore, Pakistan: Ilmi Kitab Khana.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). New York, USA: John Wiley & Sons.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). Nonparametric statistical methods (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). Melbourne, Australia: OTexts.

James, T. M., Benson, P. G., & Sincich, T. (2010). *Statistics for business and economics* (11th ed.). USA: Prentice Hall.

Levin, R. I., & Rubin, D. S. (2009). *Statistics for management* (7th ed.). Delhi, India: Dorling Kindersley Ltd.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2005). *Statistical techniques in business and economics* (12th ed.). USA: McGraw-Hill Irwin.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and applications (3rd ed.). New York, USA: John Wiley & Sons.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50–60.

Swed, F. S., & Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. The Annals of Mathematical Statistics, 14(1), 66–87.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability and statistics for engineers and scientists* (9th ed.). Boston, MA: Pearson.

APPENDIX: STATISTICAL TABLES

TABLE 1: STANDARD NORMAL DISTRIBUTION

Cumulative Probabilities $P(Z \leq z)$

The table gives the area under the standard normal curve to the left of z, i.e. $P(Z \leq z)$. Entries are given for $z = 0.00\, to\, z = 3.49$. For negative z, use the symmetry property: $P(Z \leq -z) = 1 - P(Z \leq z)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	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	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
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
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

TABLE 2: STUDENT'S t-DISTRIBUTION

Upper-Tail Critical Values $t(\alpha, \nu)$

The table gives $t(\alpha, \nu)$ such that $P(t > t(\alpha, \nu)) = \alpha$ (one-tailed). For a two-tailed test at level α*, the critical value is $t(\alpha/2, \nu)$. The header shows the one-tailed α; the two-tailed equivalent 2α is shown in parentheses.

N	0.100 (0.200)	0.050 (0.100)	0.025 (0.050)	0.010 (0.020)	0.005 (0.010)	0.0025 (0.005)	0.001 (0.002)	0.0005 (0.001)
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
Inf	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

TABLE 3: CHI-SQUARE (χ^2) DISTRIBUTION

Critical Values $\chi^2(P,\nu)$ where $P(\chi^2 \leq x) = P$

The table gives $\chi^2(P,\nu)$ such that $P(\chi^2 \leq \chi^2(P,\nu)) = P$. For a right-tailed test at level α, use the column $P = 1 - \alpha$. For a left-tailed test at level α, use the column $P = \alpha$.

N	P=0.005	P=0.010	P=0.025	P=0.050	P=0.100	P=0.950	P=0.975	P=0.990	P=0.995
1	0.000	0.000	0.001	0.004	0.016	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	124.342	129.561	135.807	140.169

TABLE 4: WILCOXON SIGNED-RANK TEST

Critical Values of W (One-Tailed α)

The table gives the critical value W such that $P(W^+ \leq W) \leq \alpha$ under H_0. For a two-tailed test at level 2α, reject H_0 if $min(W^+, W^-) \leq W$. For a one-tailed test at level α, reject H_0 if $W^+ \leq W$ (when the alternative is that the median is below zero) or $W^- \leq W$ (above zero).

N	α = 0.10	α = 0.05	α = 0.02	α = 0.01
5	0	0	0	0
6	2	0	0	0
7	3	2	0	0
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68

TABLE 5: SPEARMAN RANK CORRELATION COEFFICIENT

Critical Values of $|r_s|$ for $H_0: \rho_s = 0$ (Two-Tailed)

Reject $H_0: \rho_s = 0$ if $|r_s| \geq$ the tabled critical value. Based on the t-approximation: $t = r_s\sqrt{[(n-2)/(1-r_s^2)]}$ with $n-2$ degrees of freedom.

N	α = 0.10	α = 0.05	α = 0.02	α = 0.01
5	0.8054	0.8783	0.9343	0.9587
6	0.7293	0.8114	0.8822	0.9172
7	0.6694	0.7545	0.8329	0.8745
8	0.6215	0.7067	0.7887	0.8343
9	0.5822	0.6664	0.7498	0.7977
10	0.5494	0.6319	0.7155	0.7646
11	0.5214	0.6021	0.6851	0.7348
12	0.4973	0.5760	0.6581	0.7079
13	0.4762	0.5529	0.6339	0.6835
14	0.4575	0.5324	0.6120	0.6614
15	0.4409	0.5140	0.5923	0.6411
16	0.4259	0.4973	0.5742	0.6226
17	0.4124	0.4821	0.5577	0.6055
18	0.4000	0.4683	0.5425	0.5897
19	0.3887	0.4555	0.5285	0.5751
20	0.3783	0.4438	0.5155	0.5614
21	0.3687	0.4329	0.5034	0.5487
22	0.3598	0.4227	0.4921	0.5368
23	0.3515	0.4132	0.4815	0.5256
24	0.3438	0.4044	0.4716	0.5151
25	0.3365	0.3961	0.4622	0.5052
26	0.3297	0.3882	0.4534	0.4958
27	0.3233	0.3809	0.4451	0.4869
28	0.3172	0.3739	0.4372	0.4785
29	0.3115	0.3673	0.4297	0.4705
30	0.3061	0.3610	0.4226	0.4629

TABLE 6: KOLMOGOROV-SMIRNOV TEST

Critical Values of D_n (One-Sample Test)

The table gives $D_n(\alpha)$ such that $P(D_n > D_n(\alpha)) = \alpha$ under H_0. Reject H_0 (that the sample comes from F_0) if the observed statistic $D_n = sup_x |F_n(x) - F_0(x)|$ exceeds the tabled value.

N	α = 0.200	α = 0.100	α = 0.050	α = 0.025	α = 0.010
1	1.0730	1.2240	1.3580	1.4800	1.6280
2	0.7587	0.8655	0.9603	1.0465	1.1512
3	0.6195	0.7067	0.7840	0.8545	0.9399
4	0.5365	0.6120	0.6790	0.7400	0.8140
5	0.4799	0.5474	0.6073	0.6619	0.7281
6	0.4381	0.4997	0.5544	0.6042	0.6646
7	0.4056	0.4626	0.5133	0.5594	0.6153
8	0.3794	0.4327	0.4801	0.5233	0.5756
9	0.3577	0.4080	0.4527	0.4933	0.5427
10	0.3393	0.3871	0.4294	0.4680	0.5148
11	0.3235	0.3690	0.4095	0.4462	0.4909
12	0.3097	0.3533	0.3920	0.4272	0.4700
13	0.2976	0.3395	0.3766	0.4105	0.4515
14	0.2868	0.3271	0.3629	0.3955	0.4351
15	0.2770	0.3160	0.3506	0.3821	0.4203
16	0.2682	0.3060	0.3395	0.3700	0.4070
17	0.2602	0.2969	0.3294	0.3590	0.3948
18	0.2529	0.2885	0.3201	0.3488	0.3837
19	0.2462	0.2808	0.3115	0.3395	0.3735
20	0.2399	0.2737	0.3037	0.3309	0.3640
21	0.2341	0.2671	0.2963	0.3230	0.3553
22	0.2288	0.2610	0.2895	0.3155	0.3471
23	0.2237	0.2552	0.2832	0.3086	0.3395
24	0.2190	0.2498	0.2772	0.3021	0.3323
25	0.2146	0.2448	0.2716	0.2960	0.3256
26	0.2104	0.2400	0.2663	0.2903	0.3193
27	0.2065	0.2356	0.2613	0.2848	0.3133
28	0.2028	0.2313	0.2566	0.2797	0.3077
29	0.1993	0.2273	0.2522	0.2748	0.3023
30	0.1959	0.2235	0.2479	0.2702	0.2972
31	0.1927	0.2198	0.2439	0.2658	0.2924
32	0.1897	0.2164	0.2401	0.2616	0.2878
33	0.1868	0.2131	0.2364	0.2576	0.2834
34	0.1840	0.2099	0.2329	0.2538	0.2792
35	0.1814	0.2069	0.2295	0.2502	0.2752
40	0.1697	0.1935	0.2147	0.2340	0.2574
45	0.1600	0.1825	0.2024	0.2206	0.2427
50	0.1517	0.1731	0.1921	0.2093	0.2302
n > 50	1.073/√n	1.224/√n	1.358/√n	1.480/√n	1.628/√n

TABLE 7: PEARSON CORRELATION COEFFICIENT

Critical Values of $|r|$ for H_0: $\rho = 0$ (Two-Tailed)

Reject H_0: $\rho = 0$ if $|r| \geq$ the tabled critical value. Here $v = n - 2$ is the degrees of freedom. Critical values use the t-approximation: $t = r\sqrt{[(n-2)/(1-r^2)]}$.

v = n−2	α=0.10	α=0.05	α=0.02	α=0.01	α=0.002	α=0.001
1	0.8545	0.9969	0.9995	0.9999	1.0000	1.0000
2	0.7583	0.9500	0.9800	0.9900	0.9980	0.9990
3	0.6886	0.8783	0.9343	0.9587	0.9859	0.9911
4	0.6352	0.8114	0.8822	0.9172	0.9633	0.9741
5	0.5926	0.7545	0.8329	0.8745	0.9350	0.9509
6	0.5575	0.7067	0.7888	0.8343	0.9049	0.9249
7	0.5280	0.6664	0.7498	0.7976	0.8751	0.8983
8	0.5027	0.6319	0.7154	0.7646	0.8467	0.8721
9	0.4808	0.6020	0.6850	0.7348	0.8199	0.8471
10	0.4615	0.5760	0.6581	0.7079	0.7950	0.8233
11	0.4443	0.5529	0.6339	0.6836	0.7717	0.8010
12	0.4289	0.5324	0.6120	0.6614	0.7502	0.7800
13	0.4151	0.5139	0.5922	0.6411	0.7301	0.7604
14	0.4024	0.4973	0.5742	0.6226	0.7114	0.7419
15	0.3909	0.4821	0.5577	0.6055	0.6940	0.7247
16	0.3803	0.4683	0.5425	0.5897	0.6777	0.7084
17	0.3705	0.4556	0.5285	0.5750	0.6624	0.6932
18	0.3615	0.4438	0.5154	0.5614	0.6480	0.6788
19	0.3531	0.4329	0.5033	0.5487	0.6346	0.6652
20	0.3452	0.4227	0.4921	0.5368	0.6219	0.6524
25	0.3125	0.3809	0.4451	0.4869	0.5679	0.5974
30	0.2876	0.3493	0.4093	0.4487	0.5257	0.5541
40	0.2517	0.3044	0.3578	0.3931	0.4634	0.4896
50	0.2266	0.2733	0.3218	0.3542	0.4188	0.4432
60	0.2077	0.2500	0.2948	0.3248	0.3851	0.4078
80	0.1809	0.2172	0.2565	0.2830	0.3364	0.3568
100	0.1623	0.1946	0.2301	0.2540	0.3025	0.3211

TABLE 8: F-DISTRIBUTION CRITICAL VALUES ($\alpha = 0.05$)

Values of F such that $P(F > F\alpha, df_1, df_2) = 0.05$

Rows: df_2 (denominator degrees of freedom). Columns: df_1 (numerator degrees of freedom). Reject H_0 if your computed $F - ratio$ exceeds the tabled value.

df_2	df_1																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161.4	199.5	215.7	224.5	230.1	233.9	236.7	238.8	240.5	241.8	243.9	245.9	248.0	249.0	250.1	251.1	252.2	253.2
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	19.45	19.46	19.47	19.48	19.49
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	8.64	8.62	8.59	8.57	8.55
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	5.77	5.75	5.72	5.69	5.66
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	4.53	4.50	4.46	4.43	4.40
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	3.84	3.81	3.77	3.74	3.70
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	3.41	3.38	3.34	3.30	3.27
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	3.12	3.08	3.04	3.01	2.97
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	2.90	2.86	2.83	2.79	2.75
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	2.74	2.70	2.66	2.62	2.58
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	2.61	2.57	2.53	2.49	2.45
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	2.51	2.47	2.43	2.38	2.34
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	2.42	2.38	2.34	2.30	2.25
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	2.35	2.31	2.27	2.22	2.18
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	2.29	2.25	2.20	2.16	2.11
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	2.24	2.19	2.15	2.11	2.06
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	2.19	2.15	2.10	2.06	2.01
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	2.15	2.11	2.06	2.02	1.97
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	2.11	2.07	2.03	1.98	1.93
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	2.08	2.04	1.99	1.95	1.90
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	2.03	1.98	1.94	1.89	1.84
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	1.98	1.94	1.89	1.84	1.79
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	1.95	1.90	1.85	1.80	1.75
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	1.91	1.87	1.82	1.77	1.71
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	1.89	1.84	1.79	1.74	1.68
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	1.79	1.74	1.69	1.64	1.58
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	1.70	1.65	1.59	1.53	1.47
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	1.61	1.55	1.50	1.43	1.35

TABLE 9: F-DISTRIBUTION CRITICAL VALUES ($\alpha = 0.01$)

Values of F such that $P(F > F\alpha, df_1, df_2) = 0.01$

Rows: df_2 (denominator degrees of freedom). Columns: df_1 (numerator degrees of freedom). Reject H_0 if your computed $F - ratio$ exceeds the tabled value.

df_2	df_1																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	4052.1	4999.5	5403.3	5624.5	5763.6	5858.9	5928.3	5981.0	6022.4	6055.8	6106.3	6157.2	6208.7	6234.6	6260.6	6286.7	6313.0	6339.3
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53

TABLE 10: MANN-WHITNEY U TEST CRITICAL VALUES

Lower Critical Values of U (Two-Tailed, $\alpha = 0.05$)

The table gives the critical value U_0. For a two-tailed test at $\alpha = 0.05$, reject H_0 if $min(U, n_1n_2 - U) \leq U_0$. For a one-tailed test at $\alpha = 0.05$, use the table with the appropriate group labelling.

$n_1 \setminus n_2$	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	1	2	2
4	0	0	0	0	1	2	3	3	4	5
5	0	0	0	1	2	3	4	6	7	8
6	0	0	0	2	3	5	6	8	9	11
7	0	0	1	3	4	6	8	10	12	14
8	0	0	1	3	6	8	10	12	15	17
9	0	0	2	4	7	9	12	15	17	20
10	0	0	2	5	8	11	14	17	20	23