
Contents

List of Figures	xv
List of Tables	xxv
Preface to the Second Edition	xxix
Preface to the First Edition	xxxii
1 What Is R?	1
1.1 Introduction to R	1
1.2 Downloading and Installing R	1
1.2.1 Installing R under Windows	2
1.2.2 Launching R	4
1.2.3 A First Look at R (Interactive Mode)	4
1.3 Vectors	6
1.3.1 Naming Cautions	10
1.3.2 Vector Indexing	10
1.3.3 Generating Vector Sequences and Repeating Vector Constants	11
1.3.4 Filtering Vectors	12
1.4 Mode and Class of an Object	13
1.5 Getting Help	14
1.6 External Editors	15
1.7 RStudio	16
1.8 Packages	19
1.9 R Data Structures	21
1.9.1 Arrays and Matrices	21
1.9.2 Vector and Matrix Operations	27
1.9.3 Factors	28
1.9.4 Lists	29
1.9.5 Data Frames	31
1.9.5.1 Creating Data Frames	32
1.9.5.2 Accessing Data Frames	33
1.9.5.3 Accessing Data from Packages	37
1.10 Reading and Saving Data in R	39
1.10.1 Using <code>read.table()</code>	40
1.10.2 Using <code>download.file()</code>	41
1.10.3 Reading Data from Secure Websites	42
1.10.4 Using <code>scan()</code>	44
1.10.5 Reading Excel (<code>.xlsx</code>) Files	45
1.10.6 Saving Data Frames to External Files	47
1.11 Working with Data	47
1.11.1 Dealing with NA Values	51
1.11.2 Creating New Variables in a Data Frame	54

1.11.3	Sorting a Data Frame by One or More of Its Columns	55
1.11.4	Merging Data Frames	56
1.12	Using Logical Operators with Data Frames	58
1.13	Tables	62
1.14	Summarizing Functions	66
1.15	Probability Functions	68
1.16	Flow Control	69
1.17	Creating Functions	74
1.18	Simple Imputation	78
1.19	Using <code>plot()</code>	80
1.20	Coordinate Systems and Traditional Graphic's States	85
1.21	Problems	91
2	Exploring Data	97
2.1	What Is Statistics?	97
2.2	Data	97
2.3	Displaying Qualitative Data	98
2.3.1	Tables	98
2.3.2	Barplots	100
2.3.3	Dot Charts	100
2.3.4	Pie Charts	103
2.4	Displaying Quantitative Data	104
2.4.1	Stem-and-Leaf Plots	104
2.4.2	Strip Charts	105
2.4.3	Density Curves for Exploring Univariate Data	107
2.4.3.1	Histograms	107
2.4.3.2	Kernel Density Estimators	114
2.5	Summary Measures of Location	120
2.5.1	The Mean	120
2.5.2	The Median	121
2.5.3	Mode	123
2.5.4	Quantiles	124
2.5.5	Hinges and the Five-Number Summary	126
2.5.6	Boxplots	127
2.6	Summary Measures of Spread	129
2.6.1	Range	130
2.6.2	Interquartile Range	130
2.6.3	Variance	131
2.6.4	Sample Coefficient of Variation	132
2.6.5	The Median Absolute Deviation (<i>MAD</i>)	133
2.7	Bivariate Data	134
2.7.1	Two-Way Contingency Tables	134
2.7.2	Graphical Representations of Two-Way Contingency Tables	136
2.7.3	Comparing Samples	138
2.7.4	Relationships between Two Numeric Variables	143
2.7.5	Correlation	144
2.7.6	Fitting Lines to Bivariate Data	147
2.8	Complex Plot Arrangements	151
2.9	Multivariate Data	154
2.9.1	Graphs for Categorical Data	156
2.9.2	Lattice Graphs	162

2.9.3	Arranging Several Lattice Graphs on a Single Page	165
2.9.4	Panel Functions	167
2.9.5	Graphics with <code>ggplot2</code>	169
2.9.5.1	Shading a Region of a Density Curve	177
2.9.5.2	Violin Plots	180
2.9.5.3	Adding a Smoothed Line	184
2.9.5.4	Choropleth Maps	188
2.9.6	Arranging Several <code>ggplot</code> Graphs on a Single Page	191
2.10	Problems	193
3	General Probability and Random Variables	199
3.1	Introduction	199
3.2	Counting Techniques	199
3.2.1	Sampling with Replacement	199
3.2.2	Sampling without Replacement	200
3.2.3	Combinations	201
3.3	Axiomatic Probability	202
3.3.1	Sample Space and Events	203
3.3.2	Set Theory	203
3.3.3	Interpreting Probability	204
3.3.3.1	Relative Frequency Approach to Probability	204
3.3.3.2	Axiomatic Approach to Probability	205
3.3.4	Conditional Probability	208
3.3.5	The Law of Total Probability and Bayes' Rule	210
3.3.6	Independent Events	213
3.4	Random Variables	214
3.4.1	Discrete Random Variables	215
3.4.1.1	Mode, Median, and Percentiles	217
3.4.1.2	Expected Values	217
3.4.1.3	Moments	219
3.4.2	Continuous Random Variables	222
3.4.2.1	Numerical Integration with R	225
3.4.2.2	Mode, Median, and Percentiles	226
3.4.2.3	Expected Values	229
3.4.3	Markov's Theorem and Chebyshev's Inequality	231
3.4.4	Weak Law of Large Numbers	233
3.4.5	Skewness	234
3.5	Moment Generating Functions	235
3.6	Problems	238
4	Univariate Probability Distributions	249
4.1	Introduction	249
4.2	Discrete Univariate Distributions	249
4.2.1	Discrete Uniform Distribution	249
4.2.2	Bernoulli and Binomial Distributions	250
4.2.3	Poisson Distribution	256
4.2.4	Geometric Distribution	263
4.2.5	Negative Binomial Distribution	266
4.2.6	Hypergeometric Distribution	269
4.3	Continuous Univariate Distributions	271
4.3.1	Uniform Distribution (Continuous)	272

4.3.2	Exponential Distribution	276
4.3.3	Gamma Distribution	282
4.3.4	Hazard Function, Reliability Function, and Failure Rate	286
4.3.5	Weibull Distribution	291
4.3.6	Beta Distribution	293
4.3.7	Normal (Gaussian) Distribution	296
4.4	Problems	307
5	Multivariate Probability Distributions	315
5.1	Joint Distribution of Two Random Variables	315
5.1.1	Joint pdf for Two Discrete Random Variables	315
5.1.2	Joint pdf for Two Continuous Random Variables	317
5.2	Independent Random Variables	319
5.3	Several Random Variables	319
5.4	Conditional Distributions	322
5.5	Expected Values, Covariance, and Correlation	326
5.5.1	Expected Values	326
5.5.2	Covariance	327
5.5.3	Correlation	330
5.6	Multinomial Distribution	332
5.7	Bivariate Normal Distribution	333
5.8	Problems	342
6	Sampling and Sampling Distributions	351
6.1	Sampling	351
6.1.1	Simple Random Sampling	352
6.1.2	Stratified Sampling	354
6.1.3	Systematic Sampling	355
6.1.4	Cluster Sampling	355
6.2	Parameters	356
6.2.1	Infinite Populations' Parameters	356
6.2.2	Finite Populations' Parameters	357
6.3	Estimators	357
6.3.1	Plug-In Principle	359
6.4	Sampling Distribution of the Sample Mean	359
6.5	Sampling Distribution for a Statistic from an Infinite Population	367
6.5.1	Sampling Distribution for the Sample Mean	367
6.5.1.1	First Case: Sampling Distribution of \bar{X} When Sampling from a Normal Distribution	368
6.5.1.2	Second Case: Sampling Distribution of \bar{X} When X Is Not a Normal Random Variable	370
6.5.2	Sampling Distribution for $\bar{X} - \bar{Y}$ When Sampling from Two Independent Normal Populations	375
6.5.3	Sampling Distribution for the Sample Proportion	377
6.5.4	Expected Value and Variance of the Uncorrected Sample Variance and the Sample Variance	382
6.6	Sampling Distributions Associated with the Normal Distribution	383
6.6.1	Chi-Square Distribution (χ^2)	383
6.6.1.1	The Relationship between the χ^2 Distribution and the Normal Distribution	386

6.6.1.2	Sampling Distribution for S_u^2 and S^2 When Sampling from Normal Populations	389
6.6.2	t -Distribution	394
6.6.3	The F Distribution	397
6.7	Problems	400
7	Point Estimation	405
7.1	Introduction	405
7.2	Properties of Point Estimators	405
7.2.1	Mean Square Error	405
7.2.2	Unbiased Estimators	406
7.2.3	Efficiency	409
7.2.3.1	Relative Efficiency	410
7.2.4	Consistent Estimators	414
7.2.5	Robust Estimators	415
7.3	Point Estimation Techniques	416
7.3.1	Method of Moments Estimators	417
7.3.2	Likelihood and Maximum Likelihood Estimators	419
7.3.2.1	Fisher Information	431
7.3.2.2	Fisher Information for Several Parameters	433
7.3.2.3	Properties of Maximum Likelihood Estimators	435
7.3.2.4	Finding Maximum Likelihood Estimators for Multiple Parameters	440
7.3.2.5	Multi-Parameter Properties of MLEs	442
7.4	Problems	444
8	Confidence Intervals	453
8.1	Introduction	453
8.2	Confidence Intervals for Population Means	454
8.2.1	Confidence Interval for the Population Mean When Sampling from a Normal Distribution with Known Population Variance	454
8.2.1.1	Determining Required Sample Size	460
8.2.2	Confidence Interval for the Population Mean When Sampling from a Normal Distribution with Unknown Population Variance	464
8.2.3	Confidence Interval for the Difference in Population Means When Sampling from Independent Normal Distributions with Known Equal Variances	466
8.2.4	Confidence Interval for the Difference in Population Means When Sampling from Independent Normal Distributions with Known but Unequal Variances	470
8.2.5	Confidence Interval for the Difference in Means When Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal	474
8.2.6	Confidence Interval for a Difference in Means When Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal	476
8.2.7	Confidence Interval for the Mean Difference When the Differences Have a Normal Distribution	480
8.3	Confidence Intervals for Population Variances	482
8.3.1	Confidence Interval for the Population Variance When Sampling from a Normal Population	482

8.3.2	Confidence Interval for the Ratio of Population Variances When Sampling from Independent Normal Distributions	487
8.4	Confidence Intervals Based on Large Samples	490
8.4.1	Confidence Interval for the Population Proportion	491
8.4.1.1	Score Confidence Interval	496
8.4.1.2	Agresti-Coull Confidence Interval for the Population Proportion	498
8.4.1.3	Clopper-Pearson Interval for the Population Proportion	498
8.4.1.4	So Which Confidence Interval Do I Use?	498
8.4.2	Confidence Interval for a Difference in Population Proportions	506
8.4.3	Confidence Interval for the Mean of a Poisson Random Variable	508
8.5	Problems	510
9	Hypothesis Testing	519
9.1	Introduction	519
9.2	Type I and Type II Errors	520
9.3	Power Function	524
9.4	Uniformly Most Powerful Test	527
9.5	ϕ -Value or Critical Level	529
9.6	Tests of Significance	530
9.7	Hypothesis Tests for Population Means	532
9.7.1	Test for the Population Mean When Sampling from a Normal Distribution with Known Population Variance	532
9.7.2	Test for the Population Mean When Sampling from a Normal Distribution with Unknown Population Variance	535
9.7.3	Test for the Difference in Population Means When Sampling from Independent Normal Distributions with Known Variances	542
9.7.4	Test for the Difference in Means When Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal	544
9.7.5	Test for a Difference in Means When Sampling from Independent Normal Distributions with Variances That Are Unknown and Not Assumed Equal	548
9.7.6	Test for the Mean Difference When the Differences Have a Normal Distribution	551
9.8	Hypothesis Tests for Population Variances	555
9.8.1	Test for the Population Variance When Sampling from a Normal Distribution	555
9.8.2	Test for Equality of Variances When Sampling from Independent Normal Distributions	558
9.9	Hypothesis Tests for Population Proportions	562
9.9.1	Testing the Proportion of Successes in a Binomial Experiment (Exact Test)	562
9.9.2	Testing the Proportion of Successes in a Binomial Experiment (Normal Approximation)	565
9.9.3	Testing Equality of Proportions with Fisher's Exact Test	569
9.9.4	Large Sample Approximation for Testing the Difference of Two Proportions	574
9.10	Problems	579

10 Nonparametric Methods	587
10.1 Introduction	587
10.2 Sign Test	588
10.2.1 Confidence Interval Based on the Sign Test	588
10.2.2 Normal Approximation to the Sign Test	589
10.3 Wilcoxon Signed-Rank Test	594
10.3.1 Confidence Interval for ψ Based on the Wilcoxon Signed-Rank Test	599
10.3.2 Normal Approximation to the Wilcoxon Signed-Rank Test	603
10.4 The Wilcoxon Rank-Sum or the Mann-Whitney U -Test	608
10.4.1 Confidence Interval Based on the Mann-Whitney U -Test	612
10.4.2 Normal Approximation to the Wilcoxon Rank-Sum and Mann-Whitney U -Tests	615
10.5 The Kruskal-Wallis Test	622
10.6 Friedman Test for Randomized Block Designs	629
10.7 Goodness-of-Fit Tests	634
10.7.1 The Chi-Square Goodness-of-Fit Test	635
10.7.2 Kolmogorov-Smirnov Goodness-of-Fit Test	640
10.7.3 Shapiro-Wilk Normality Test	647
10.8 Categorical Data Analysis	649
10.8.1 Test of Independence	651
10.8.2 Test of Homogeneity	653
10.9 Nonparametric Bootstrapping	656
10.9.1 Bootstrap Paradigm	656
10.9.2 Confidence Intervals	665
10.9.3 Bootstrapping and Regression	677
10.10 Permutation Tests	681
10.11 Problems	688
11 Experimental Design	697
11.1 Introduction	697
11.2 Fixed Effects Model	702
11.3 Analysis of Variance (ANOVA) for the One-Way Fixed Effects Model	703
11.4 Power and the Non-Central F Distribution	709
11.5 Checking Assumptions	718
11.5.1 Checking for Independence of Errors	719
11.5.2 Checking for Normality of Errors	720
11.5.3 Checking for Constant Variance	722
11.6 Fixing Problems	724
11.6.1 Non-Normality	725
11.6.2 Non-Constant Variance	726
11.7 Multiple Comparisons of Means	730
11.7.1 Fisher's Least Significant Difference	731
11.7.2 The Tukey's Honestly Significant Difference	732
11.7.3 Displaying Pairwise Comparisons	733
11.8 Other Comparisons among the Means	733
11.8.1 Orthogonal Contrasts	734
11.8.2 The Scheffé Method for All Contrasts	740
11.9 Summary of Comparisons of Means	740
11.10 Random Effects Model (Variance Components Model)	745
11.11 Randomized Complete Block Design	748
11.12 Two-Factor Factorial Design	760

11.13	Problems	771
12	Regression	781
12.1	Introduction	781
12.2	Simple Linear Regression	783
12.3	Multiple Linear Regression	784
12.4	Ordinary Least Squares	785
12.5	Properties of the Fitted Regression Line	788
12.6	Using Matrix Notation with Ordinary Least Squares	789
12.7	The Method of Maximum Likelihood	796
12.8	The Sampling Distribution of $\hat{\beta}$	797
12.9	ANOVA Approach to Regression	800
12.9.1	ANOVA with Simple Linear Regression	801
12.9.2	ANOVA with Multiple Linear Regression	805
12.9.3	Coefficient of Determination	806
12.9.4	Extra Sum of Squares	807
12.9.4.1	Tests on a Single Parameter	812
12.9.4.2	Tests on Subsets of the Regression Parameters	815
12.10	General Linear Hypothesis	816
12.11	Model Building	822
12.11.1	Testing-Based Procedures	822
12.11.1.1	Backward Elimination	822
12.11.1.2	Forward Selection	822
12.11.1.3	Stepwise Regression	822
12.11.1.4	Criterion-Based Procedures	829
12.11.1.5	Summary	834
12.11.2	Diagnostics	834
12.11.2.1	Checking Error Assumptions	834
12.11.2.1.1	Assessing Normality and Constant Variance	836
12.11.2.1.2	Testing Autocorrelation	836
12.11.2.2	Identifying Unusual Observations	838
12.11.2.3	High Leverage Observations	844
12.11.3	Transformations	850
12.11.3.1	Collinearity	853
12.11.3.2	Transformations for Non-Normality and Unequal Error Variances	856
12.12	Model Validation	862
12.12.1	The Validation Set Approach	863
12.12.2	Leave-One-Out Cross-Validation	864
12.12.3	k -Fold Cross-Validation	865
12.13	Interpreting a Logarithmically Transformed Model	871
12.14	Qualitative Predictors	873
12.15	Estimation of the Mean Response for New Values \mathbf{X}_h	880
12.16	Prediction and Sampling Distribution of New Observations $Y_{h(\text{new})}$	880
12.17	Simultaneous Confidence Intervals	883
12.17.1	Simultaneous Confidence Intervals for Several Mean Responses — Confidence Band	884
12.17.2	Predictions of g New Observations	884
12.17.3	Distinguishing Pointwise Confidence Envelopes from Confidence Bands	884
12.18	Problems	891

A R Commands	903
B Quadratic Forms and Random Vectors and Matrices	917
B.1 Quadratic Forms	917
B.2 Random Vectors and Matrices	918
B.3 Variance of Random Vectors	918
Bibliography	921
Index	931