

---

# Contents

<b>1</b>	<b>What Is R?</b>	<b>1</b>
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 . . . . .	18
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 . . . . .	31
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> . . . . .	39
1.10.2	Using <code>download.file()</code> . . . . .	40
1.10.3	Reading Data from Secure Web Sites . . . . .	41
1.10.4	Using <code>scan()</code> . . . . .	43
1.10.5	Reading Excel ( <code>.xlsx</code> ) Files . . . . .	44
1.10.6	Saving Data Frames to External Files . . . . .	45
1.11	Working with Data . . . . .	47
1.11.1	Dealing with NA Values . . . . .	50
1.11.2	Creating New Variables in a Data Frame . . . . .	53
1.11.3	Sorting a Data Frame by One or More of Its Columns . . . . .	54
1.11.4	Merging Data Frames . . . . .	56
1.12	Using Logical Operators with Data Frames . . . . .	57
1.13	Tables . . . . .	61
1.14	Summarizing Functions . . . . .	66
1.15	Probability Functions . . . . .	68
1.16	Flow Control . . . . .	68

1.17	Creating Functions . . . . .	73
1.18	Simple Imputation . . . . .	77
1.19	Using <code>plot()</code> . . . . .	79
1.20	Coordinate Systems and Traditional Graphic's States . . . . .	85
1.21	Problems . . . . .	90
<b>2</b>	<b>Exploring Data</b>	<b>97</b>
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	Sorting a Data Frame by One or More of Its Columns . . . . .	146
2.7.7	Fitting Lines to Bivariate Data . . . . .	147
2.8	Complex Plot Arrangements . . . . .	152
2.9	Multivariate Data . . . . .	155
2.9.1	Graphs for Categorical Data . . . . .	157
2.9.2	Lattice Graphs . . . . .	163
2.9.3	Arranging Several Lattice Graphs on a Single Page . . . . .	166
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 . . . . .	179
2.9.5.2	Violin Plots . . . . .	180
2.9.5.3	Adding a Smoothed Line . . . . .	184

2.9.5.4	Choropleth Maps . . . . .	189
2.9.6	Arranging Several <code>ggplot</code> Graphs on a Single Page . . . . .	191
2.10	Problems . . . . .	194
<b>3</b>	<b>General Probability and Random Variables</b>	<b>201</b>
3.1	Introduction . . . . .	201
3.2	Counting Techniques . . . . .	201
3.2.1	Sampling With Replacement . . . . .	201
3.2.2	Sampling Without Replacement . . . . .	202
3.2.3	Combinations . . . . .	203
3.3	Axiomatic Probability . . . . .	204
3.3.1	Sample Space and Events . . . . .	205
3.3.2	Set Theory . . . . .	205
3.3.3	Interpreting Probability . . . . .	206
3.3.3.1	Relative Frequency Approach to Probability . . . . .	206
3.3.3.2	Axiomatic Approach to Probability . . . . .	207
3.3.4	Conditional Probability . . . . .	210
3.3.5	The Law of Total Probability and Bayes' Rule . . . . .	212
3.3.6	Independent Events . . . . .	215
3.4	Random Variables . . . . .	216
3.4.1	Discrete Random Variables . . . . .	217
3.4.1.1	Mode, Median, and Percentiles . . . . .	219
3.4.1.2	Expected Values . . . . .	219
3.4.1.3	Moments . . . . .	221
3.4.2	Continuous Random Variables . . . . .	224
3.4.2.1	Numerical Integration with R . . . . .	227
3.4.2.2	Mode, Median, and Percentiles . . . . .	228
3.4.2.3	Expected Values . . . . .	231
3.4.3	Markov's Theorem and Chebyshev's Inequality . . . . .	233
3.4.4	Weak Law of Large Numbers . . . . .	235
3.4.5	Skewness . . . . .	236
3.5	Moment Generating Functions . . . . .	237
3.6	Problems . . . . .	240
<b>4</b>	<b>Univariate Probability Distributions</b>	<b>249</b>
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 . . . . .	272
4.3.1	Uniform Distribution (Continuous) . . . . .	272
4.3.2	Exponential Distribution . . . . .	277
4.3.3	Gamma Distribution . . . . .	282
4.3.4	Hazard Function, Reliability Function, and Failure Rate . . . . .	287
4.3.5	Weibull Distribution . . . . .	291
4.3.6	Beta Distribution . . . . .	293
4.3.7	Normal (Gaussian) Distribution . . . . .	296

4.4	Problems	308
<b>5</b>	<b>Multivariate Probability Distributions</b>	<b>317</b>
5.1	Joint Distribution of Two Random Variables	317
5.1.1	Joint pdf for Two Discrete Random Variables	317
5.1.2	Joint pdf for Two Continuous Random Variables	319
5.2	Independent Random Variables	320
5.3	Several Random Variables	321
5.4	Conditional Distributions	324
5.5	Expected Values, Covariance, and Correlation	328
5.5.1	Expected Values	328
5.5.2	Covariance	330
5.5.3	Correlation	333
5.6	Multinomial Distribution	334
5.7	Bivariate Normal Distribution	336
5.8	Problems	344
<b>6</b>	<b>Sampling and Sampling Distributions</b>	<b>353</b>
6.1	Sampling	353
6.1.1	Simple Random Sampling	354
6.1.2	Stratified Sampling	356
6.1.3	Systematic Sampling	357
6.1.4	Cluster Sampling	357
6.2	Parameters	358
6.2.1	Infinite Populations' Parameters	358
6.2.2	Finite Populations' Parameters	359
6.3	Estimators	359
6.3.1	Plug-In Principle	361
6.4	Sampling Distribution of the Sample Mean	361
6.5	Sampling Distribution for a Statistic from an Infinite Population	369
6.5.1	Sampling Distribution for the Sample Mean	369
6.5.1.1	First Case: Sampling Distribution of $\bar{X}$ when Sampling from a Normal Distribution	370
6.5.1.2	Second Case: Sampling Distribution of $\bar{X}$ when $X$ Is not a Normal Random Variable	372
6.5.2	Sampling Distribution for $\bar{X} - \bar{Y}$ when Sampling from Two Independent Normal Populations	377
6.5.3	Sampling Distribution for the Sample Proportion	379
6.5.4	Expected Value and Variance of the Uncorrected Sample Variance and the Sample Variance	383
6.6	Sampling Distributions Associated with the Normal Distribution	385
6.6.1	Chi-Square Distribution ( $\chi^2$ )	385
6.6.1.1	The Relationship between the $\chi^2$ Distribution and the Normal Distribution	388
6.6.1.2	Sampling Distribution for $S_u^2$ and $S^2$ when Sampling from Normal Populations	390
6.6.2	$t$ -Distribution	396
6.6.3	The $F$ Distribution	399
6.7	Problems	402

<b>7</b>	<b>Point Estimation</b>	<b>409</b>
7.1	Introduction	409
7.2	Properties of Point Estimators	409
7.2.1	Mean Square Error	409
7.2.2	Unbiased Estimators	410
7.2.3	Efficiency	413
7.2.3.1	Relative Efficiency	414
7.2.4	Consistent Estimators	417
7.2.5	Robust Estimators	419
7.3	Point Estimation Techniques	420
7.3.1	Method of Moments Estimators	421
7.3.2	Likelihood and Maximum Likelihood Estimators	423
7.3.2.1	Fisher Information	435
7.3.2.2	Fisher Information for Several Parameters	437
7.3.2.3	Properties of Maximum Likelihood Estimators	439
7.3.2.4	Finding Maximum Likelihood Estimators for Multiple Parameters	444
7.3.2.5	Multi-Parameter Properties of MLEs	447
7.4	Problems	448
<b>8</b>	<b>Confidence Intervals</b>	<b>457</b>
8.1	Introduction	457
8.2	Confidence Intervals for Population Means	458
8.2.1	Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Known Population Variance	458
8.2.1.1	Determining Required Sample Size	464
8.2.2	Confidence Interval for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance	468
8.2.3	Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Equal Variances	470
8.2.4	Confidence Interval for the Difference in Population Means when Sampling from Independent Normal Distributions with Known but Unequal Variances	474
8.2.5	Confidence Interval for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal	478
8.2.6	Confidence Interval for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and Unequal	480
8.2.7	Confidence Interval for the Mean Difference when the Differences Have a Normal Distribution	484
8.3	Confidence Intervals for Population Variances	488
8.3.1	Confidence Interval for the Population Variance of a Normal Population	488
8.3.2	Confidence Interval for the Ratio of Population Variances when Sampling from Independent Normal Distributions	492
8.4	Confidence Intervals Based on Large Samples	495
8.4.1	Confidence Interval for the Population Proportion	496
8.4.1.1	Score Confidence Interval	501

8.4.1.2	Agresti-Coull Confidence Interval for the Population Proportion . . . . .	502
8.4.1.3	The Clopper-Pearson Interval for the Population Proportion . . . . .	503
8.4.1.4	So Which Confidence Interval Do I Use? . . . . .	503
8.4.2	Confidence Interval for a Difference in Population Proportions . .	511
8.4.3	Confidence Interval for the Mean of a Poisson Random Variable .	513
8.5	Problems . . . . .	515
<b>9</b>	<b>Hypothesis Testing</b>	<b>523</b>
9.1	Introduction . . . . .	523
9.2	Type I and Type II Errors . . . . .	524
9.3	Power Function . . . . .	528
9.4	Uniformly Most Powerful Test . . . . .	531
9.5	$p$ -Value or Critical Level . . . . .	533
9.6	Tests of Significance . . . . .	534
9.7	Hypothesis Tests for Population Means . . . . .	536
9.7.1	Test for the Population Mean when Sampling from a Normal Distribution with Known Population Variance . . . . .	536
9.7.2	Test for the Population Mean when Sampling from a Normal Distribution with Unknown Population Variance . . . . .	539
9.7.3	Test for the Difference in Population Means when Sampling from Independent Normal Distributions with Known Variances . . . . .	546
9.7.4	Test for the Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown but Assumed Equal . . . . .	548
9.7.5	Test for a Difference in Means when Sampling from Independent Normal Distributions with Variances That Are Unknown and not Assumed Equal . . . . .	552
9.7.6	Test for the Mean Difference when the Differences Have a Normal Distribution . . . . .	555
9.8	Hypothesis Tests for Population Variances . . . . .	559
9.8.1	Test for the Population Variance when Sampling from a Normal Distribution . . . . .	559
9.8.2	Test for Equality of Variances when Sampling from Independent Normal Distributions . . . . .	562
9.9	Hypothesis Tests for Population Proportions . . . . .	566
9.9.1	Testing the Proportion of Successes in a Binomial Experiment (Exact Test) . . . . .	566
9.9.2	Testing the Proportion of Successes in a Binomial Experiment (Normal Approximation) . . . . .	569
9.9.3	Testing Equality of Proportions with Fisher's Exact Test . . . . .	573
9.9.4	Large Sample Approximation for Testing the Difference of Two Proportions . . . . .	579
9.10	Problems . . . . .	583
<b>10</b>	<b>Nonparametric Methods</b>	<b>591</b>
10.1	Introduction . . . . .	591
10.2	Sign Test . . . . .	592
10.2.1	Confidence Interval Based on the Sign Test . . . . .	592
10.2.2	Normal Approximation to the Sign Test . . . . .	593

10.3	Wilcoxon Signed-Rank Test . . . . .	598
10.3.1	Confidence Interval for $\psi$ Based on the Wilcoxon Signed-Rank Test . . . . .	603
10.3.2	Normal Approximation to the Wilcoxon Signed-Rank Test . . . . .	607
10.4	The Wilcoxon Rank-Sum or the Mann-Whitney $U$ -Test . . . . .	612
10.4.1	Confidence Interval Based on the Mann-Whitney $U$ -Test . . . . .	616
10.4.2	Normal Approximation to the Wilcoxon Rank-Sum and Mann-Whitney $U$ -Tests . . . . .	619
10.5	The Kruskal-Wallis Test . . . . .	626
10.6	Friedman Test for Randomized Block Designs . . . . .	633
10.7	Goodness-of-Fit Tests . . . . .	638
10.7.1	The Chi-Square Goodness-of-Fit Test . . . . .	639
10.7.2	Kolmogorov-Smirnov Goodness-of-Fit Test . . . . .	645
10.7.3	Shapiro-Wilk Normality Test . . . . .	652
10.8	Categorical Data Analysis . . . . .	653
10.8.1	Test of Independence . . . . .	655
10.8.2	Test of Homogeneity . . . . .	658
10.9	Nonparametric Bootstrapping . . . . .	660
10.9.1	Bootstrap Paradigm . . . . .	660
10.9.2	Confidence Intervals . . . . .	670
10.9.3	Bootstrapping and Regression . . . . .	680
10.10	Permutation Tests . . . . .	685
10.11	Problems . . . . .	692
<b>11</b>	<b>Experimental Design</b>	<b>701</b>
11.1	Introduction . . . . .	701
11.2	Fixed Effects Model . . . . .	706
11.3	Analysis of Variance (ANOVA) for the One-Way Fixed Effects Model . . . . .	707
11.4	Power and the Non-Central $F$ Distribution . . . . .	713
11.5	Checking Assumptions . . . . .	722
11.5.1	Checking for Independence of Errors . . . . .	723
11.5.2	Checking for Normality of Errors . . . . .	724
11.5.3	Checking for Constant Variance . . . . .	726
11.6	Fixing Problems . . . . .	728
11.6.1	Non-Normality . . . . .	729
11.6.2	Non-Constant Variance . . . . .	730
11.7	Multiple Comparisons of Means . . . . .	734
11.7.1	Fisher's Least Significant Difference . . . . .	734
11.7.2	The Tukey's Honestly Significant Difference . . . . .	736
11.7.3	Displaying Pairwise Comparisons . . . . .	737
11.8	Other Comparisons among the Means . . . . .	737
11.8.1	Orthogonal Contrasts . . . . .	738
11.8.2	The Scheffé Method for All Contrasts . . . . .	744
11.9	Summary of Comparisons of Means . . . . .	744
11.10	Random Effects Model (Variance Components Model) . . . . .	749
11.11	Randomized Complete Block Design . . . . .	752
11.12	Two-Factor Factorial Design . . . . .	764
11.13	Problems . . . . .	775

<b>12 Regression</b>	<b>785</b>
12.1 Introduction	785
12.2 Simple Linear Regression	787
12.3 Multiple Linear Regression	788
12.4 Ordinary Least Squares	789
12.5 Properties of the Fitted Regression Line	792
12.6 Using Matrix Notation with Ordinary Least Squares	793
12.7 The Method of Maximum Likelihood	800
12.8 The Sampling Distribution of $\hat{\beta}$	801
12.9 ANOVA Approach to Regression	804
12.9.1 ANOVA with Simple Linear Regression	805
12.9.2 ANOVA with Multiple Linear Regression	809
12.9.3 Coefficient of Determination	810
12.9.4 Extra Sum of Squares	811
12.9.4.1 Tests on a Single Parameter	816
12.9.4.2 Tests on Subsets of the Regression Parameters	819
12.10 General Linear Hypothesis	820
12.11 Model Building	826
12.11.1 Testing-Based Procedures	826
12.11.1.1 Backward Elimination	826
12.11.1.2 Forward Selection	826
12.11.1.3 Stepwise Regression	826
12.11.1.4 Criterion-Based Procedures	833
12.11.1.5 Summary	838
12.11.2 Diagnostics	838
12.11.2.1 Checking Error Assumptions	838
12.11.2.1.1 Assessing Normality and Constant Variance	840
12.11.2.1.2 Testing Autocorrelation	840
12.11.2.2 Identifying Unusual Observations	842
12.11.2.3 High Leverage Observations	848
12.11.3 Transformations	854
12.11.3.1 Collinearity	857
12.11.3.2 Transformations for Non-Normality and Unequal Error Variances	860
12.12 Model Validation	866
12.12.1 The Validation Set Approach	867
12.12.2 Leave-One-Out Cross-Validation	868
12.12.3 $k$ -Fold Cross-Validation	869
12.13 Interpreting a Logarithmically Transformed Model	875
12.14 Qualitative Predictors	877
12.15 Estimation of the Mean Response for New Values $\mathbf{X}_h$	883
12.16 Prediction and Sampling Distribution of New Observations $Y_{h(\text{new})}$	884
12.17 Simultaneous Confidence Intervals	887
12.17.1 Simultaneous Confidence Intervals for Several Mean Responses — Confidence Band	888
12.17.2 Predictions of $g$ New Observations	888
12.17.3 Distinguishing Pointwise Confidence Envelopes from Confidence Bands	888
12.18 Problems	895
<b>A R Commands</b>	<b>907</b>



<b>B Quadratic Forms and Random Vectors and Matrices</b>	<b>921</b>
B.1 Quadratic Forms . . . . .	921
B.2 Random Vectors and Matrices . . . . .	922
B.3 Variance of Random Vectors . . . . .	922
<b>References</b>	<b>925</b>
<b>Index</b>	<b>935</b>