

Contents

- 1 Introduction 1**
 - 1.1 Statistical Methods in the Context of Scientific Studies 1
 - 1.2 Sampling 3
 - 1.3 Observational Studies and Experiments 4
 - 1.4 Data Exploration and Analysis 5
 - 1.5 Statistical Inference 5
 - 1.6 Computation 6
 - 1.6.1 Using R-Commander 6
 - 1.6.2 Using R 10
 - 1.7 Advanced 11
 - 1.7.1 More on Sampling 11
 - 1.7.2 More on Observational Studies 12
 - 1.7.3 More on Experiments 13
 - 1.7.4 Cross-Sectional, Longitudinal, and Time Series Data 14
 - 1.8 Exercises 15
- 2 Data Exploration 17**
 - 2.1 Data Visualization and Summary Statistics 17
 - 2.2 Variable Types 17
 - 2.3 Exploring Categorical Variables 20
 - 2.3.1 Relative Frequency and Percentage 21
 - 2.3.2 Bar Graph 22
 - 2.3.3 Pie Chart 24
 - 2.4 Exploring Numerical Variables 25
 - 2.4.1 Histograms 26
 - 2.4.2 Mean and Median 32
 - 2.4.3 Variance and Standard Deviation 34
 - 2.4.4 Quantiles 37
 - 2.4.5 Boxplots 38
 - 2.5 Data Preprocessing 40
 - 2.5.1 Missing Data 40

2.5.2	Outliers	41
2.5.3	Data Transformation	43
2.5.4	Creating New Variable Based on Two or More Existing Variables	44
2.5.5	Creating Categories for Numerical Variables	45
2.6	Advanced	46
2.6.1	Coefficient of Variation	46
2.6.2	Scaling and Shifting Variables	48
2.6.3	Variable Standardization	50
2.6.4	Data Exploration with R Programming	51
2.7	Exercises	57
3	Exploring Relationships	61
3.1	Visualizing and Summarizing Relationships Between Variables	61
3.2	Relationships Between Two Numerical Random Variables	61
3.3	Relationships Between Categorical Variables	69
3.4	Relationships Between Numerical and Categorical Variables	73
3.5	Advanced	76
3.6	Exercises	79
4	Probability	83
4.1	Probability as a Measure of Uncertainty	83
4.2	Some Commonly Used Genetic Terms	83
4.3	The Sample Space	84
4.4	Probability Measure	86
4.5	Complement, Union, and Intersection	87
4.5.1	Complement	89
4.5.2	Union	90
4.5.3	Intersection	91
4.5.4	Joint vs. Marginal Probability	91
4.6	Disjoint Events	92
4.7	Conditional Probabilities	93
4.8	The Law of Total Probability	95
4.9	Independent Events	97
4.10	Bayes' Theorem	98
4.10.1	Application of Bayes' Theorem in Medical Diagnosis	99
4.10.2	Bayesian Statistics	101
4.11	Interpretation of Probability as the Relative Frequency	101
4.12	Advanced	102
4.12.1	Using Tree Diagrams to Obtain Joint Probabilities	103
4.12.2	Making Decisions under Uncertainty	105
4.13	Exercises	107
5	Random Variables and Probability Distributions	109
5.1	Random Variables	109
5.2	Discrete vs. Continuous	110
5.3	Probability Distributions	111

5.4	Discrete Probability Distributions	112
5.4.1	Bernoulli Distribution	113
5.4.2	Binomial Distribution	115
5.4.3	Poisson Distribution	119
5.5	Continuous Probability Distributions	121
5.5.1	Probability Density Curves and Density Histograms	124
5.5.2	Normal Distribution	125
5.5.3	Student's t -distribution	130
5.6	Cumulative Distribution Function and Quantiles	131
5.7	Scaling and Shifting Random Variables	134
5.8	Sum of Two Random Variables	135
5.9	Advanced	137
5.9.1	More on Probability Distributions	137
5.9.2	Some Other Commonly Used Probability Distributions	138
5.9.3	Quantile–Quantile Plots	142
5.9.4	Probability Distributions with R Programming	143
5.10	Exercises	148
6	Estimation	151
6.1	Parameter Estimation	151
6.2	Point Estimation	152
6.2.1	Population Mean	152
6.2.2	Population Variance	154
6.3	Sampling Distribution	156
6.4	Confidence Intervals for the Population Mean	158
6.5	Confidence Interval When the Population Variance Is Unknown	162
6.6	Using Central Limit Theorem for Confidence Interval	163
6.7	Confidence Intervals for the Population Proportion	166
6.8	Margin of Error	167
6.9	Advanced	168
6.9.1	Deriving Confidence Intervals	168
6.9.2	Sample Size Estimation	169
6.10	Exercises	170
7	Hypothesis Testing	173
7.1	Introduction	173
7.2	Hypothesis Testing for the Population Mean	174
7.3	Statistical Significance	176
7.3.1	z -Tests of the Population Mean	177
7.3.2	Interpretation of p -value	178
7.3.3	One-Sided Hypothesis Testing	181
7.3.4	Two-Sided Hypothesis Testing	183
7.4	Hypothesis Testing Using t -tests	184
7.5	Hypothesis Testing for Population Proportion	186
7.6	Advanced	188
7.6.1	Test of Normality	188

7.6.2	Hypothesis Testing with R Programming	188
7.7	Exercises	190
8	Statistical Inference for the Relationship Between Two Variables . .	193
8.1	Introduction	193
8.2	Relationship Between a Numerical Variable and a Binary Variable .	193
8.2.1	Two-Sample t -tests for Comparing the Means	197
8.2.2	Pooled t -test	203
8.2.3	Paired t -test	203
8.3	Inference about the Relationship Between Two Binary Variables . .	208
8.4	Inference Regarding the Linear Relationship Between Two Numerical Variables	211
8.5	Advanced	215
8.5.1	Two-Sample t -test Using R	216
8.5.2	Correlation Test Using R	217
8.6	Exercises	218
9	Analysis of Variance (ANOVA)	221
9.1	Introduction	221
9.2	Analysis of Variance (ANOVA)	221
9.3	The Assumptions of ANOVA	228
9.4	Advanced	230
9.4.1	Two-Way ANOVA	230
9.4.2	ANOVA Using R	232
9.5	Exercises	233
10	Analysis of Categorical Variables	235
10.1	Introduction	235
10.2	Pearson's χ^2 Test for One Categorical Variable	236
10.2.1	Binary Variables	236
10.2.2	Categorical Variables with Multiple Categories	239
10.3	Pearson's χ^2 Test of Independence	240
10.4	Entering Contingency Tables into R-Commander	245
10.5	Advanced	246
10.5.1	Fisher's Exact Test	246
10.5.2	Pearson's χ^2 Test Using R	248
10.6	Exercises	250
11	Regression Analysis	253
11.1	Introduction	253
11.2	Linear Regression Models with One Binary Explanatory Variable .	254
11.3	Statistical Inference Using Simple Linear Regression Models . . .	259
11.3.1	Confidence Interval for Regression Coefficients	261
11.3.2	Hypothesis Testing with Simple Linear Regression Models	263
11.4	Linear Regression Models with One Numerical Explanatory Variable	264

11.5	Goodness of Fit	270
11.6	Model Assumptions and Diagnostics	272
11.7	Multiple Linear Regression	275
11.8	Advanced	282
11.8.1	Interaction	282
11.8.2	Linear Regression Models in R	285
11.9	Exercises	288
12	Clustering	291
12.1	Introduction	291
12.2	K-means Clustering	293
12.3	Hierarchical Clustering	295
12.4	Advanced	297
12.4.1	Standardizing Variables Before Clustering	297
12.4.2	Clustering in R	298
12.5	Exercises	301
13	Bayesian Analysis	303
13.1	Introduction	303
13.2	A Simple Case of Bayesian Analysis for Population Proportion	303
13.3	Prior and Posterior Probabilities	305
13.4	The General Form of Bayesian Analysis for Population Proportion	306
13.5	Bayesian Inference	310
13.5.1	Estimation	310
13.5.2	Hypothesis Testing	311
13.6	Advanced	313
13.7	Exercises	314
Appendix A	Installing R and R-Commander	317
A.1	Installing R	317
A.2	Installing R-Commander	317
A.2.1	From the Command Line	317
A.2.2	From the Menu Bar	318
A.3	Starting R-Commander	319
Appendix B	Basic R	323
B.1	Starting with R	323
B.2	Creating Objects in R	324
B.3	Vectors	326
B.4	Matrices	332
B.5	Data Frames	335
B.5.1	Creating Data Frames Using a Spreadsheet-Like Environment	337
B.5.2	Importing Data from Text Files	337
B.6	Lists	339
B.7	Loading Add-on Packages	340
B.8	Conditional Statements	341

B.9 Loops 343

B.10 Creating Functions 344

References 347

Index 349

Biostatistics with R

An Introduction to Statistics Through Biological Data

Shahbaba, B.

2012, XVI, 352 p. 162 illus., 73 illus. in color., Softcover

ISBN: 978-1-4614-1301-1