

Table of Contents

Preface	xv
1 Introduction	1
1.1 What Is Biostatistics?	1
1.2 Data — The Key Component of a Study	2
1.3 Design — The Road to Relevant Data	4
1.4 Replication — Part of the Scientific Method	6
1.5 Applying Statistical Methods	7
Exercises	7
References	8
2 Data and Numbers	9
2.1 Data: Numerical Representation	9
2.2 Observations and Variables	10
2.3 Scales Used with Variables	10
2.4 Reliability and Validity	11
2.5 Randomized Response Technique	13
2.6 Common Data Problems	14
Conclusion	17
Exercises	18
References	20
3 Descriptive Methods	21
3.1 Introduction to Descriptive Methods	21
3.2 Tabular and Graphical Presentation of Data	22
3.2.1 Frequency Tables	23
3.2.2 Line Graphs	24
3.2.3 Bar Charts	27
3.2.4 Histograms	30
3.2.5 Stem-and-Leaf Plots	35
3.2.6 Dot Plots	37
3.2.7 Scatter Plots	38
3.3 Measures of Central Tendency	39
3.3.1 Mean, Median, and Mode	40
3.3.2 Use of the Measures of Central Tendency	42
3.3.3 The Geometric Mean	42

3.4	Measures of Variability	45
3.4.1	Range and Percentiles	45
3.4.2	Box Plots	47
3.4.3	Variance and Standard Deviation	48
3.5	Rates and Ratios	51
3.5.1	Crude and Specific Rates	52
3.5.2	Adjusted Rates	53
3.6	Measures of Change over Time	55
3.6.1	Linear Growth	55
3.6.2	Geometric Growth	57
3.6.3	Exponential Growth	58
3.7	Correlation Coefficients	60
3.7.1	Pearson Correlation Coefficient	60
3.7.2	Spearman Rank Correlation Coefficient	63
	Conclusion	64
	Exercises	64
	References	68
4	Probability and Life Tables	71
4.1	A Definition of Probability	71
4.2	Rules for Calculating Probabilities	73
4.2.1	Addition Rule for Probabilities	73
4.2.2	Conditional Probabilities	75
4.2.3	Independent Events	77
4.3	Definitions from Epidemiology	80
4.3.1	Rates and Probabilities	80
4.3.2	Sensitivity, Specificity, and Predicted Value Positive and Negative	81
4.3.3	Receiver Operating Characteristic Plot	83
4.4	Bayes' Theorem	84
4.5	Probability in Sampling	87
4.5.1	Sampling with Replacement	87
4.5.2	Sampling without Replacement	88
4.6	Estimating Probabilities by Simulation	89
4.7	Probability and the Life Table	91
4.7.1	The First Four Columns in the Life Table	93
4.7.2	Some Uses of the Life Table	95
4.7.3	Expected Values in the Life Table	96
4.7.4	Other Expected Values in the Life Table	98
	Conclusion	99
	Exercises	99
	References	102
5	Probability Distributions	103
5.1	The Binomial Distribution	103
5.1.1	Binomial Probabilities	103
5.1.2	Mean and Variance of the Binomial Distribution	108
5.1.3	Shapes of the Binomial Distribution	110

5.2	The Poisson Distribution	111
5.2.1	Poisson Probabilities	111
5.2.2	Mean and Variance of the Poisson Distribution	113
5.2.3	Finding Poisson Probabilities	114
5.3	The Normal Distribution	116
5.3.1	Normal Probabilities	116
5.3.2	Transforming to the Standard Normal Distribution	118
5.3.3	Calculation of Normal Probabilities	119
5.3.4	The Normal Probability Plot	122
5.4	The Central Limit Theorem	124
5.5	Approximations to the Binomial and Poisson Distributions	126
5.5.1	Normal Approximation to the Binomial Distribution	126
5.5.2	Normal Approximation to the Poisson Distribution	129
	Conclusion	131
	Exercises	131
	References	133
6	Study Designs	135
6.1	Design: Putting Chance to Work	135
6.2	Sample Surveys and Experiments	137
6.3	Sampling and Sample Designs	138
6.3.1	Sampling Frame	139
6.3.2	Importance of Probability Sampling	140
6.3.3	Simple Random Sampling	141
6.3.4	Systematic Sampling	142
6.3.5	Stratified Random Sampling	144
6.3.6	Cluster Sampling	144
6.3.7	Problems Due to Unintended Sampling	145
6.4	Designed Experiments	148
6.4.1	Comparison Groups and Randomization	149
6.4.2	Random Assignment	150
6.4.3	Sample Size	152
6.4.4	Single- and Double-Blind Experiments	154
6.4.5	Blocking and Extraneous Variables	155
6.4.6	Limitations of Experiments	156
6.5	Variations in Study Designs	158
6.5.1	The Crossover Design	158
6.5.2	The Case Control Design	159
6.5.3	The Cohort Study Design	160
	Conclusion	160
	Exercises	161
	References	166
7	Interval Estimation	169
7.1	Prediction, Confidence, and Tolerance Intervals	169
7.2	Distribution-Free Intervals	170
7.2.1	Prediction Interval	170
7.2.2	Confidence Interval	171
7.2.3	Tolerance Interval	175

7.3	Confidence Intervals Based on the Normal Distribution	176
7.3.1	Confidence Interval for the Mean	177
7.3.2	Confidence Interval for a Proportion	182
7.3.3	Confidence Interval for Crude and Adjusted Rates	185
7.4	Confidence Interval for the Difference of Two Means and Proportions	188
7.4.1	Difference of Two Independent Means	188
7.4.2	Difference of Two Dependent Means	194
7.4.3	Difference of Two Independent Proportions	196
7.4.4	Difference of Two Dependent Proportions	197
7.5	Confidence Interval and Sample Size	198
7.6	Confidence Intervals for Other Measures	200
7.6.1	Confidence Interval for the Variance	201
7.6.2	Confidence Interval for Pearson Correlation Coefficient	203
7.7	Prediction and Tolerance Intervals Based on the Normal Distribution	205
7.7.1	Prediction Interval	205
7.7.2	Tolerance Interval	206
	Conclusion	206
	Exercises	207
	References	211
8	Tests of Hypotheses	213
8.1	Preliminaries in Tests of Hypotheses	213
8.1.1	Terms Used in Hypothesis Testing	215
8.1.2	Determination of the Decision Rule	216
8.1.3	Relationship of the Decision Rule, α and β	218
8.1.4	Conducting the Test	221
8.2	Testing Hypotheses about the Mean	223
8.2.1	Known Variance	223
8.2.2	Unknown Variance	228
8.3	Testing Hypotheses about the Proportion and Rates	229
8.4	Testing Hypotheses about the Variance	231
8.5	Testing Hypotheses about the Pearson Correlation Coefficient	232
8.6	Testing Hypotheses about the Difference of Two Means	234
8.6.1	Difference of Two Independent Means	234
8.6.2	Difference of Two Dependent Means	237
8.7	Testing Hypotheses about the Difference of Two Proportions	238
8.7.1	Difference of Two Independent Proportions	238
8.7.2	Difference of Two Dependent Proportions	239
8.8	Tests of Hypotheses and Sample Size	240
8.9	Statistical and Practical Significance	243
	Conclusion	243
	Exercises	244
	References	248
9	Nonparametric Tests	249
9.1	Why Nonparametric Tests?	249
9.2	The Sign Test	249
9.3	The Wilcoxon Signed Rank Test	253

9.4	The Wilcoxon Rank Sum Test	257
9.5	The Kruskal-Wallis Test	261
9.6	The Friedman Test	262
	Conclusion	264
	Exercises	264
	References	268
10	Analysis of Categorical Data	269
10.1	The Goodness-of-Fit Test	269
10.2	The 2 by 2 Contingency Table	273
10.2.1	Comparing Two Independent Binomial Proportions	274
10.2.2	Expected Cell Counts Assuming No Association: Chi-Square Test	274
10.2.3	The Odds Ratio — a Measure of Association	277
10.2.4	Fisher's Exact Test	279
10.2.5	The Analysis of Matched-Pairs Studies	280
10.3	The r by c Contingency Table	282
10.3.1	Testing Hypothesis of No Association	282
10.3.2	Testing Hypothesis of No Trend	284
10.4	Multiple 2 by 2 Contingency Tables	286
10.4.1	Analyzing the Tables Separately	287
10.4.2	The Cochran-Mantel-Haenszel Test	288
10.4.3	The Mantel-Haenszel Common Odds Ratio	290
	Conclusion	291
	Exercises	291
	References	295
11	Analysis of Survival Data	297
11.1	Data Collection in Follow-up Studies	297
11.2	The Life-Table Method	299
11.3	The Product-Limit Method	306
11.4	Comparison of Two Survival Distributions	310
11.4.1	The CMH Test	310
11.4.2	The Normal Distribution Approach	313
11.4.3	The Log-Rank Test	313
11.4.4	Use of the CMH Approach with Small Data Sets	314
	Conclusion	316
	Exercises	316
	References	320
12	Analysis of Variance	323
12.1	Assumptions for Use of the ANOVA	323
12.2	One-Way ANOVA	324
12.2.1	Sums of Squares and Mean Squares	325
12.2.2	The F Statistic	326
12.2.3	The ANOVA Table	327
12.3	Multiple Comparisons	329

12.3.1	Error Rates: Individual and Family	329
12.3.2	The Tukey-Kramer Method	330
12.3.3	Fisher's Least Significant Difference Method	330
12.3.4	Dunnett's Method	331
12.4	Two-Way ANOVA for the Randomized Block Design with m Replicates	332
12.5	Two-Way ANOVA with Interaction	335
12.6	Linear Model Representation of the ANOVA	339
12.6.1	The Completely Randomized Design	339
12.6.2	The Randomized Block Design with m Replicates	341
12.6.3	Two-Way ANOVA with Interaction	341
12.7	ANOVA with Unequal Numbers of Observations in Subgroups	342
	Conclusion	345
	Exercises	345
	References	347
13	Linear Regression	349
13.1	Simple Linear Regression	349
13.1.1	Estimation of Coefficients	351
13.1.2	The Variance of $Y X$	353
13.1.3	The Coefficient of Determination (R^2)	355
13.2	Inference about the Coefficients	357
13.2.1	Assumptions for Inference in Linear Regression	357
13.2.2	Regression Diagnostics	358
13.2.3	The Slope Coefficient	361
13.2.4	The Y-intercept	362
13.2.5	ANOVA Table Summary	363
13.3	Interval Estimation for $\mu_{Y X}$ and $Y X$	364
13.3.1	Confidence Interval for $\mu_{Y X}$	364
13.3.2	Prediction Interval for $Y X$	366
13.4	Multiple Linear Regression	368
13.4.1	The Multiple Linear Regression Model	368
13.4.2	Specification of a Multiple Linear Regression Model	369
13.4.3	Parameter Estimates, ANOVA, and Diagnostics	374
13.4.4	Multicollinearity Problems	376
13.4.5	Extending the Regression Model: Dummy Variables	378
	Conclusion	380
	Exercises	380
	References	385
14	Logistic and Proportional Hazards Regression	387
14.1	Simple Logistic Regression	387
14.1.1	Proportion, Odds and Logit	389
14.1.2	Estimation of Parameters	390
14.1.3	Computer Output	391
14.1.4	Statistical Inference	392
14.2	Multiple Logistic Regression	394
14.2.1	Model and Assumptions	394
14.2.2	Residuals	398

14.2.3	Goodness-of-Fit Statistics	399
14.2.4	The ROC Curve	401
14.3	Ordered Logistic Regression	403
14.4	Conditional Logistic Regression	407
14.5	Introduction to Proportional Hazard Regression	409
	Conclusion	415
	Exercises	416
	References	419
15	Analysis of Survey Data	421
15.1	Introduction to Design-Based Inference	421
15.2	Components in Design-Based Analysis	422
15.2.1	Sample Weights	422
15.2.2	Poststratification	423
15.2.3	The Design Effect	424
15.3	Strategies for Variance Estimation	426
15.3.1	Replicated Sampling: A General Approach	426
15.3.2	Balanced Repeated Replication	427
15.3.3	Jackknife Repeated Replication	428
15.3.4	Linearization Method	430
15.4	Strategies for Analysis	431
15.4.1	Preliminary Analysis	432
15.4.2	Subpopulation Analysis	433
15.5	Some Analytic Examples	434
15.5.1	Descriptive Analysis	434
15.5.2	Contingency Table Analysis	435
15.5.3	Linear and Logistic Regression Analysis	437
	Conclusion	440
	Exercises	440
	References	442
	Appendix A: Review of Basic Mathematic Concepts	445
	Appendix B: Statistical Tables	451
B1.	Random Digits	452
B2.	Binomial Probabilities	454
B3.	Poisson Probabilities	458
B4.	Cumulative Distribution Function for Standard Normal Distribution	461
B5.	Critical Values for the t Distribution	463
B6.	Graphs for Binomial Confidence Interval	464
B7.	Critical Values for the Chi-Square (χ^2) Distribution	466
B8.	Factors, k , for Two-Sided Tolerance Limits for Normal Distributions	467
B9.	Critical Values for Wilcoxon Signed Rank Test	469
B10.	Critical Values for Wilcoxon Rank Sum Test	470

B11. Critical Values for the F Distribution	474
B12. Upper Percentage Points of the Studentized Range	477
B13. t for Comparisons Between p Treatment Means and a Control for a Joint Confidence Coefficient of $p = 0.95$ and $p = 0.99$	479
Appendix C: Selected Governmental Sources of Biostatistical Data	481
CI. Population Census Data	481
CII. Vital Statistics	482
CIII. Sample Surveys	483
CIV. Life Tables	484
Appendix D: Solutions to Selected Exercises	487
Index	493