

Contents

Preface	vii
1 Audience	vii
2 Structure	viii
3 Data and Programs	ix
4 Software	x
4.1 Microsoft Windows	xi
4.2 Unix	xi
5 Exercises	xi
1 Introduction and Motivation	1
1.1 Statistics in Context	3
1.2 Examples of Uses of Statistics	4
1.2.1 Investigation of Salary Discrimination	4
1.2.2 Measuring Body Fat	5
1.2.3 Minimizing Film Thickness	5
1.2.4 Surveys	6
1.2.5 Bringing Pharmaceutical Products to Market	6
1.3 The Rest of the Book	7
1.3.1 Fundamentals	7
1.3.2 Linear Models	7
1.3.3 Other Techniques	9
1.3.4 New Graphical Display Techniques	9
2 Data and Statistics	11
2.1 Types of Data	11
2.2 Data Display and Calculation	12
2.2.1 Presentation	13

2.2.2	Rounding	13
2.3	Importing Data	14
2.3.1	S-PLUS	14
2.3.2	SAS	15
2.3.3	Data Rearrangement	15
2.4	Analysis with Missing Data	16
2.4.1	Missing Data in S-PLUS	16
2.4.2	Missing Data in SAS	17
2.5	Tables and Graphs	17
2.6	Files for <i>Statistical Analysis and Data Display</i> (HH)	18
2.6.1	Datasets	18
2.6.2	Code, Transcripts, and Figures	18
2.6.3	Functions and Macros	19
2.6.4	Software	19
3	Statistics Concepts	21
3.1	A Brief Introduction to Probability	21
3.2	Random Variables and Probability Distributions	22
3.2.1	Discrete Versus Continuous Probability Distributions	23
3.2.2	Displaying Probability Distributions	24
3.3	Concepts That Are Used When Discussing Distributions	27
3.3.1	Expectation and Variance of Random Variables	27
3.3.2	Median of Random Variables	28
3.3.3	Symmetric and Skewed Distributions	28
3.3.4	Displays of Univariate Data	30
3.3.5	Multivariate Distributions—Covariance and Correlation	34
3.4	Three Probability Distributions	37
3.4.1	The Binomial Distribution	37
3.4.2	The Normal Distribution	38
3.4.3	The (Student's) t Distribution	39
3.5	Sampling Distributions	40
3.6	Estimation	41
3.6.1	Statistical Models	41
3.6.2	Point and Interval Estimators	42
3.6.3	Criteria for Point Estimators	42
3.6.4	Confidence Interval Estimation	43
3.6.5	Example—Confidence Interval on the Mean μ of a Population Having Known Standard Deviation	44
3.6.6	Example—One-Sided Confidence Intervals	44
3.7	Hypothesis Testing	45
3.8	Examples of Statistical Tests	47
3.9	Power and Operating Characteristic (O.C.) Curves	49
3.10	Sampling	52
3.10.1	Simple Random Sampling	53

3.10.2	Stratified Random Sampling	53
3.10.3	Cluster Random Sampling	54
3.10.4	Systematic Random Sampling	55
3.10.5	Standard Errors of Sample Means	56
3.10.6	Sources of Bias in Samples	56
3.11	Exercises	57
4	Graphs	63
4.1	Definition	64
4.2	Example—Ecological Correlation	64
4.3	Scatterplots	65
4.4	Scatterplot Matrix	67
4.5	Example—Life Expectancy	71
4.6	Scatterplot Matrices—Continued	74
4.7	Data Transformations	78
4.8	Life Expectancy Example—Continued	82
4.9	SAS Graphics	85
4.10	Exercises	87
5	Introductory Inference	91
5.1	Normal (z) Intervals and Tests	91
5.1.1	Test of a Hypothesis Concerning the Mean of a Population Having Known Standard Deviation	92
5.1.2	Confidence Intervals for Unknown Population Proportion p	93
5.1.3	Tests on an Unknown Population Proportion p	94
5.1.4	Example—One-Sided Hypothesis Test Concerning a Population Proportion	94
5.2	t -intervals and Tests for the Mean of a Population Having Unknown Standard Deviation	95
5.3	Confidence Interval on the Variance or Standard Deviation of a Normal Population	96
5.4	Comparisons of Two Populations Based on Independent Samples	97
5.4.1	Confidence Intervals on the Difference Between Two Population Proportions	98
5.4.2	Confidence Interval on the Difference of Between Two Means	98
5.4.3	Tests Comparing Two Population Means When the Samples Are Independent	99
5.4.4	Comparing the Variances of Two Normal Populations	100
5.5	Paired Data	101
5.6	Sample Size Determination	105
5.6.1	Sample Size for Estimation	105
5.6.2	Sample Size for Hypothesis Testing	106
5.7	Goodness of Fit	106
5.7.1	Chi-square Goodness-of-Fit Test	107

5.7.2	Example—Test of Goodness-of-Fit to a Discrete Uniform Distribution	108
5.7.3	Example—Test of Goodness-of-Fit to a Binomial Distribution	108
5.8	Normal Probability Plots and Quantile Plots	110
5.8.1	Normal Probability Plots	112
5.8.2	Example—Comparing t -Distributions	113
5.9	Kolmogorov–Smirnov Goodness-of-Fit Tests	114
5.10	Maximum Likelihood	117
5.10.1	Maximum Likelihood Estimation	118
5.10.2	Likelihood Ratio Tests	119
5.11	Exercises	119
6	One-Way Analysis of Variance	123
6.1	Example—Catalyst Data	123
6.2	Fixed Effects	127
6.3	Multiple Comparisons—Tukey Procedure for Comparing All Pairs of Means .	130
6.4	Random Effects	135
6.5	Expected Mean Squares (EMS)	135
6.6	Example—Catalyst Data—Continued	136
6.7	Example—Batch Data	137
6.8	Example—Turkey Data	137
6.8.1	Analysis	139
6.8.2	Interpretation	143
6.8.3	Specification of Analysis	143
6.9	Contrasts	144
6.9.1	Mathematics of Contrasts	144
6.9.2	Scaling	146
6.10	Tests of Homogeneity of Variance	147
6.11	Exercises	148
6.A	Appendix: Computation for the Analysis of Variance	153
6.A.1	Computing Notes	153
6.A.2	Computation	153
7	Multiple Comparisons	155
7.1	Multiple Comparison Procedures	156
7.1.1	Bonferroni Method	156
7.1.2	Tukey Procedure for All Pairwise Comparisons	157
7.1.3	The Dunnett Procedure for Comparing One Mean with All Others .	157
7.1.4	Simultaneously Comparing All Possible Contrasts—Scheffé and Extended Tukey	162
7.2	The Mean–Mean Multiple Comparisons Display (MMC Plot)	168
7.2.1	Difficulties with Standard Displays	168
7.2.2	Hsu and Peruggia’s Mean–Mean Scatterplot	173
7.2.3	Extensions of the Mean–Mean Display to Arbitrary Contrasts	178

7.2.4	Display of an Orthogonal Basis Set of Contrasts	180
7.2.5	Hsu and Peruggia's Pulmonary Example	182
7.3	Exercises	184
8	Linear Regression by Least Squares	187
8.1	Introduction	187
8.2	Example—Body Fat Data	188
8.3	Simple Linear Regression	190
8.3.1	Algebra	190
8.3.2	Normal Distribution Theory	192
8.3.3	Calculations	193
8.3.4	Residual Mean Square in Regression Printout	199
8.3.5	New Observations	199
8.4	Diagnostics	205
8.5	Graphics	209
8.6	Exercises	210
8.A	Appendix: Computation for Regression Analysis	213
8.A.1	S-PLUS Functions	213
8.A.2	SAS Macros and Procs	213
9	Multiple Regression—More Than One Predictor	215
9.1	Regression with Two Predictors—Least-Squares Geometry	215
9.2	Multiple Regression—Algebra	217
9.2.1	The Hat Matrix and Leverage	220
9.3	Multiple Regression—Two- X Analysis	221
9.4	Geometry of Multiple Regression	223
9.5	Programming	223
9.5.1	Model Specification	223
9.5.2	Printout Idiosyncrasies	224
9.6	Example—Albuquerque Home Price Data	225
9.7	Partial F -Tests	228
9.8	Polynomial Models	230
9.9	Models Without a Constant Term	233
9.10	Prediction	235
9.11	Example—Longley Data	236
9.12	Collinearity	241
9.13	Variable Selection	243
9.13.1	Manual Use of the Stepwise Philosophy	244
9.13.2	Automated Stepwise Regression	247
9.13.3	Automated Stepwise Modeling of the Longley Data	250
9.14	Residual Plots	254
9.14.1	Partial Residuals	254
9.14.2	Partial Residual Plots	256
9.14.3	Partial Correlation	256

9.14.4	Added Variable Plots	256
9.14.5	Interpretation of Residual Plots	257
9.15	Example—U.S. Air Pollution Data	259
9.16	Exercises	264
10	Multiple Regression—Dummy Variables and Contrasts	267
10.1	Dummy (Indicator) Variables	267
10.2	Example—Height and Weight	268
10.3	Equivalence of Linear Independent X -Variables for Regression	275
10.4	Polynomial Contrasts and Orthogonal Polynomials	277
10.4.1	Specification and Interpretation of Interaction Terms	282
10.5	Analysis Using a Concomitant Variable (Analysis of Covariance)	283
10.6	Example—Hot Dog Data	284
10.6.1	One-Way ANOVA	284
10.6.2	Concomitant Explanatory Variable	286
10.6.3	Tests of Equality of Regression Lines	292
10.7	ancova Function	294
10.8	Exercises	294
11	Multiple Regression—Regression Diagnostics	297
11.1	Example—Rent Data	297
11.1.1	Rent Levels	298
11.1.2	Alfalfa Rent Relative to Other Rent	303
11.2	Checks on Model Assumptions	309
11.2.1	Scatterplot Matrix	309
11.2.2	Residual Plots	309
11.3	Case Statistics	312
11.3.1	Leverage	315
11.3.2	Deleted Standard Deviation	316
11.3.3	Standardized and Studentized Deleted Residuals	317
11.3.4	Cook's Distance	319
11.3.5	DFFITS	321
11.3.6	DFBETAS	322
11.3.7	Calculation of Regression Diagnostics	323
11.4	Exercises	324
12	Two-Way Analysis of Variance	329
12.1	Example—Display Panel Data	329
12.2	Statistical Model	336
12.3	Main Effects and Interactions	336
12.4	Two-Way Interaction Plot	338
12.5	Sums of Squares in the Two-Way ANOVA Table	339
12.6	Treatment and Blocking Factors	339
12.7	Fixed and Random Effects	341

12.8	Randomized Complete Block Designs	342
12.9	Example—The Blood Plasma Data	344
12.10	Random Effects Models and Mixed Models	346
12.11	Introduction to Nesting	347
12.11.1	Example—Workstation Data	347
12.12	Example—Display Panel Data—Continued	349
12.13	Example—The <i>Rhizobium</i> Data	353
12.13.1	First <i>Rhizobium</i> Experiment: Alfalfa Plants	354
12.13.2	Second <i>Rhizobium</i> Experiment: Clover Plants	354
12.13.3	Initial Plots	355
12.13.4	Alfalfa Analysis	357
12.13.5	Clover Analysis	362
12.14	Models Without Interaction	371
12.15	Example—Animal Feed Data	372
12.16	Exercises	374
12.A	Appendix: Computation for the Analysis of Variance	379
13	Design of Experiments—Factorial Designs	381
13.1	A Three-Way ANOVA—Muscle Data	381
13.2	Latin Square Designs	389
13.2.1	Example—Latin Square	390
13.3	Simple Effects for Interaction Analyses	396
13.3.1	Example—The <i>filmcoat</i> Data	397
13.4	Nested Factorial Experiment	401
13.4.1	Example—Gunload Data	401
13.4.2	Example—Turkey Data (continued)	410
13.5	Specification of Model Formulas	413
13.6	Sequential and Conditional Tests	417
13.6.1	SAS Types of Sums of Squares	418
13.6.2	Example—Application to Body Fat Data	419
13.7	Exercises	421
13.A	Appendix: Orientation for Boxplots	427
14	Design of Experiments—Complex Designs	429
14.1	Confounding	429
14.2	Split Plot Designs	431
14.3	Example—Yates Oat Data	432
14.3.1	Alternate Specification	439
14.3.2	Polynomial Effects for Nitrogen	440
14.4	Introduction to Fractional Factorial Designs	442
14.4.1	Example— 2^{8-2} Design	442
14.4.2	Example— 2^{5-1} Design	444
14.5	Introduction to Crossover Designs	448
14.6	Example—Apple Tree Data	452

14.6.1	Models in Table 14.17	453
14.6.2	Multiple Comparisons	458
14.6.3	Models in Figure 14.5	460
14.7	Example— <code>testscore.dat</code>	466
14.8	The Tukey One Degree of Freedom for Nonadditivity	472
14.8.1	Example—Crash Data	472
14.8.2	Theory	481
14.9	Exercises	483
15	Bivariate Statistics—Discrete Data	487
15.1	Two-Dimensional Contingency Tables—Chi-Square Analysis	487
15.1.1	Example—Drunkenness Data	487
15.1.2	Chi-Square Analysis	490
15.2	Two-Dimensional Contingency Tables—Fisher’s Exact Test	492
15.2.1	Example—Do Juvenile Delinquents Eschew Wearing Eyeglasses?	493
15.3	Simpson’s Paradox	495
15.4	Relative Risk and Odds Ratios	498
15.4.1	Glasses (Again)	499
15.4.2	Large Sample Approximations	500
15.4.3	Example—Treating Cardiac Arrest with Therapeutic Hypothermia	500
15.5	Retrospective and Prospective Studies	503
15.6	Mantel–Haenszel Test	504
15.7	Example—Salk Polio Vaccine	506
15.8	Exercises	508
16	Nonparametrics	511
16.1	Introduction	511
16.2	Sign Test for the Location of a Single Population	512
16.3	Comparing the Locations of Paired Populations	514
16.3.1	Sign Test	514
16.3.2	Wilcoxon Signed-Ranks Test	516
16.4	Mann–Whitney Test for Two Independent Samples	520
16.5	Kruskal–Wallis Test for Comparing the Locations of at Least Three Populations	523
16.6	Exercises	526
17	Logistic Regression	527
17.1	Example—The Space Shuttle Challenger Disaster	529
17.1.1	Graphical Display	530
17.1.2	Numerical Display	533
17.2	Estimation	537
17.3	Example—Budworm Data	540
17.4	Example—Lymph Nodes	542
17.4.1	Data	542

17.4.2	Data Analysis	543
17.4.3	Additional Techniques	546
17.4.4	Diagnostics	553
17.5	Numerical Printout	553
17.6	Graphics	553
17.6.1	Conditioned Scatterplots	553
17.6.2	Scatterplot Matrix	554
17.6.3	Common Scaling in Comparable Plots	554
17.6.4	Functions of Predicted Values	555
17.7	Model Specification	556
17.7.1	S-PLUS	556
17.7.2	SAS	557
17.8	Fitting Models When the Response Is a Sample Proportion	557
17.9	LogXact	558
17.10	Exercises	558
18	Time Series Analysis	565
18.1	Introduction	565
18.2	The ARIMA Approach to Time Series Modeling	567
18.3	Autocorrelation	570
18.3.1	Autocorrelation Function (ACF)	570
18.3.2	Partial Autocorrelation Function (PACF)	570
18.4	Analysis Steps	571
18.5	Some Algebraic Development, Including Forecasting	573
18.5.1	The General ARIMA Model	573
18.5.2	Special case—The AR(1) model	574
18.5.3	Special Case—The MA(1) Model	575
18.6	Graphical Displays for Time Series Analysis	575
18.7	Models with Seasonal Components	580
18.7.1	Multiplicative Seasonal ARIMA Models	580
18.7.2	Example—co2 ARIMA(0, 1, 1) $\times$ (0, 1, 1) ₁₂ Model	581
18.7.3	Determining the Seasonal AR and MA Parameters	581
18.8	Example of a Seasonal Model—The Monthly co2 Data	582
18.8.1	Identification of the Model	582
18.8.2	Parameter Estimation and Diagnostic Checking	584
18.8.3	Forecasting	589
18.9	Exercises	589
18.A	Appendix: Graphical Displays for Time Series Analysis	618
18.A.1	Characteristics of This Presentation of the Time Series Plot	619
18.A.2	Characteristics of This Presentation of the Sample ACF and PACF Plots	619
18.A.3	Construction of Graphical Displays	620
18.A.4	User Functions Written for S-PLUS	620

A	Software	623
A.1	Statistical Software	623
A.2	Text Editing Software	624
A.2.1	Emacs	624
A.2.2	Microsoft Word	625
A.3	Word Processing Software	625
A.3.1	L ^A T _E X	626
A.3.2	Microsoft Word	626
A.4	Graphics Display Software	626
A.5	Operating Systems	627
A.6	Mathematical Fonts	627
A.7	Directory Structure	627
A.7.1	HOME Directory	627
A.7.2	HH Book Online Files	629
B	S-PLUS and R	631
B.1	Create Your Working Directory and Make the HH Library Available	632
B.1.1	Windows—Both S-PLUS and R	632
B.1.2	Windows and S-PLUS	633
B.1.3	Windows and R	634
B.1.4	Unix—Both S-PLUS and R	635
B.1.5	Unix and S-PLUS	636
B.1.6	Unix and R	636
B.2	Using S-PLUS and R with HH	637
B.3	S-PLUS for Windows—Recommended Options	638
B.4	HH Library Functions	640
B.5	Learning the S Language	640
B.6	S Language Style	643
B.7	S-PLUS Inexplicable Error Messages	645
B.8	Using S-PLUS with Emacs and ESS	647
B.9	Constructing the HH Library with S-PLUS and R	647
C	SAS	649
C.1	Make the HH Library Available	649
C.1.1	Windows	649
C.1.2	Unix	650
C.2	Using SAS with HH	652
C.2.1	Reading HH Datasets	652
C.2.2	Any Other Data Files	653
C.2.3	ASCII Data Files with TAB Characters	653
C.2.4	Windows and Unix EOL (End-of-Line) Conventions	654
C.3	Macros	655
C.4	Learning the SAS Language	655
C.5	SAS Coding Conventions	656

D	Probability Distributions	657
D.1	Common Probability Distributions with S-PLUS and SAS Commands	657
D.1.1	An Example Involving Calculations with the Binomial Distribution	661
D.2	Noncentral Probability Distributions	661
E	Editors	663
E.1	Working Style	664
E.2	Typography	665
E.3	Emacs and ESS	667
E.3.1	ESS	670
E.3.2	Mouse and Keyboard	671
E.3.3	Learning Emacs	672
E.3.4	Requirements	672
E.4	Microsoft Word	673
E.4.1	Learning Word	673
E.4.2	Requirements	673
E.5	Microsoft Excel	674
E.5.1	Database Management	674
E.5.2	Organizing Calculations	674
E.5.3	Excel as a Statistical Calculator	674
E.6	Exhortations, Some of Which Are Writing Style	677
E.6.1	Writing Style	677
E.6.2	Programming Style and Common Errors	678
E.6.3	Presentation of Results	679
F	Mathematics Preliminaries	683
F.1	Algebra Review	683
F.2	Elementary Differential Calculus	685
F.3	An Application of Differential Calculus	686
F.4	Topics in Matrix Algebra	687
F.4.1	Elementary Operations	688
F.4.2	Linear Independence	690
F.4.3	Rank	691
F.4.4	Quadratic Forms	692
F.4.5	Orthogonal Transformations	692
F.4.6	Orthogonal Basis	693
F.4.7	Matrix Factorization— QR	693
F.4.8	Matrix Factorization—Cholesky	695
F.4.9	Orthogonal Polynomials	695
F.4.10	Projection Matrices	695
F.4.11	Geometry of Matrices	695
F.4.12	Eigenvalues and Eigenvectors	696
F.4.13	Singular Value Decomposition	698
F.4.14	Generalized Inverse	698

F.4.15	Solving Linear Equations	699
F.5	Combinations and Permutations	700
F.5.1	Factorial	700
F.5.2	Permutations	700
F.5.3	Combinations	700
F.6	Exercises	701
G	Graphs Based on Cartesian Products	703
G.1	Structured Sets of Graphs	704
G.1.1	Cartesian Products	704
G.1.2	Trellis Paradigm	704
G.2	Scatterplot Matrices: <code>spiom</code> and <code>xysplom</code>	705
G.3	Cartesian Products of Sets of Functions	706
G.4	Graphs Requiring Multiple Calls to <code>xysplom</code>	706
G.5	Asymmetric Roles for the Row and Column Sets	707
G.6	Rotated Plots	707
G.7	Squared Residual Plots	708
G.8	Alternate Presentations	708
	References	709
	List of Datasets	721
	Index	723