

Contents

| | |
|--|-----|
| Preface | vii |
| 1 Wide Range of Applications | 1 |
| The Resampling Methods..... | 1 |
| Fields of Application..... | 2 |
| 2 Estimation and the Bootstrap | 7 |
| Precision of an Estimate..... | 7 |
| Stata..... | 10 |
| Applying the Bootstrap..... | 10 |
| Which Statistic Should We Use? | 10 |
| Confidence Intervals..... | 12 |
| When Variances Cannot Be Assumed to Be the Same | 12 |
| R..... | 13 |
| Stata..... | 13 |
| Testing for Equivalence..... | 14 |
| Improved Confidence Intervals | 15 |
| The Bias-Corrected Bootstrap Confidence Interval | 15 |
| Computer Code: The Bias-Corrected and -Accelerated Bootstrap | 16 |
| R..... | 16 |
| SAS..... | 17 |
| S-Plus..... | 17 |
| Stata..... | 17 |
| Balanced Bootstrap | 17 |
| Tilted Bootstrap..... | 18 |
| Block Bootstrap..... | 18 |
| Iterated Bootstrap..... | 19 |
| When the Form of the Distribution Is Known..... | 20 |
| Estimating Bias..... | 20 |
| An Example..... | 21 |
| Determining Sample Size | 22 |
| The Original Sample..... | 22 |
| The Bootstrap Sample..... | 23 |
| Summary..... | 24 |
| To Learn More | 24 |
| Exercises..... | 25 |
| 3 Software for Use with the Bootstrap and Permutation Tests | 27 |
| AFNI..... | 27 |
| Blossom Statistical Analysis Package | 27 |

| | |
|-----------------------------|----|
| Eviews..... | 28 |
| HaploView..... | 28 |
| MatLab®..... | 28 |
| NCSS..... | 28 |
| PAUP..... | 29 |
| R..... | 29 |
| SAS..... | 29 |
| S-Plus..... | 30 |
| SPSS Exact Tests..... | 30 |
| Stata..... | 30 |
| Statistical Calculator..... | 30 |
| StatXact..... | 31 |
| Testimate..... | 31 |

| | |
|---|-----------|
| 4 <u>Comparing Two Populations</u> | 33 |
| A Distribution-Free Test..... | 33 |
| A Little Math..... | 35 |
| Some Statistical Considerations..... | 35 |
| Framing the Hypothesis..... | 36 |
| Hypothesis vs. Alternative..... | 36 |
| Assumptions..... | 37 |
| General Hypotheses..... | 38 |
| Computing the p -Value..... | 39 |
| Monte Carlo..... | 39 |
| R..... | 40 |
| SPLUS..... | 40 |
| STATA..... | 40 |
| Other Two-Sample Comparisons..... | 41 |
| Two-Sided Test..... | 41 |
| Rank Tests..... | 42 |
| Matched Pairs..... | 42 |
| R Code..... | 43 |
| Stata..... | 44 |
| Test for Nonequivalence..... | 44 |
| Underlying Assumptions..... | 45 |
| Comparing Variances..... | 45 |
| R Code for Aly's Test Statistic..... | 47 |
| Unequal Sample Sizes..... | 48 |
| Preferred Method..... | 48 |
| R Code..... | 49 |
| Testing in the Presence of Nonresponders..... | 50 |
| Summary..... | 51 |
| To Learn More..... | 51 |
| Exercises..... | 52 |

| | | |
|----------|---|----|
| 5 | <u>Multiple Variables</u> | 55 |
| | Single-Valued Test Statistic | 55 |
| | Hotelling's T^2 | 55 |
| | Application to Repeated Measures | 57 |
| | The Generalized Quadratic Form | 58 |
| | Application to Epidemiology | 58 |
| | Further Generalization | 59 |
| | The MRPP Statistic | 59 |
| | Analyzing Migration Data | 60 |
| | Gene Set Enrichment Analysis | 61 |
| | Combining Univariate Tests | 62 |
| | Pesarin's Nonparametric Combination | 64 |
| | Summary | 65 |
| | To Learn More | 65 |
| | Exercises | 66 |
| 6 | <u>Experimental Design and Analysis</u> | 69 |
| | Separating Signal from Noise | 69 |
| | Blocking | 70 |
| | Analyzing a Blocked Experiment | 71 |
| | Combining Data to Obtain Improved Estimates | 71 |
| | Comparing Samples from Two Populations | 72 |
| | Randomization | 73 |
| | k -Sample Comparison | 74 |
| | Testing for Any and All Differences among Means | 74 |
| | Testing for Any and All Differences among Variances | 75 |
| | R | 76 |
| | Stata | 77 |
| | Ordered Alternatives | 77 |
| | Multiple Factors | 78 |
| | Main Effects | 79 |
| | Testing for Interactions | 81 |
| | Eliminating the Effects of Multiple Covariates | 82 |
| | Latin Squares | 83 |
| | Crossover Designs | 86 |
| | Analysis of a Complete Balanced Design | 87 |
| | Analysis of a Balanced Design When Not All Subjects Complete Treatment | 88 |
| | Which Sets of Labels Should We Rearrange? | 88 |
| | Determining Sample Size | 89 |
| | Missing Combinations | 89 |
| | Summary | 91 |
| | To Learn More | 91 |
| | Exercises | 92 |

| | |
|--|-----|
| 7 Categorical Data | 97 |
| Fisher's Exact Test | 97 |
| Computing Fisher's Exact Test | 99 |
| R | 100 |
| Two-Tailed Tests | 100 |
| Borderline Significance | 102 |
| Is the Sample Large Enough? | 103 |
| Odds Ratio | 104 |
| Stratified 2×2 s | 106 |
| Controlling the False Discovery Rate | 107 |
| Unordered $r \times c$ Contingency Tables..... | 107 |
| Test of Association | 109 |
| Causation vs. Association | 111 |
| Ordered Statistical Tables | 112 |
| Partial Dependence..... | 113 |
| Correspondence Analysis | 114 |
| More Than Two Rows and Two Columns | 114 |
| Singly Ordered Tables | 114 |
| Doubly Ordered Tables..... | 116 |
| Multidimensional Arrays | 116 |
| Summary..... | 117 |
| To Learn More | 118 |
| Exercises | 118 |
| 8 Multiple Hypotheses | 121 |
| Controlling the Family-Wise Error Rate..... | 121 |
| Microarray Analysis | 122 |
| EEG Analysis | 122 |
| Controlling the False Discovery Rate | 123 |
| Software for Performing Multiple Simultaneous Tests..... | 124 |
| AFNI | 124 |
| ExactFDR..... | 124 |
| NPCtest | 125 |
| R | 125 |
| SAS | 125 |
| Testing for Trend..... | 125 |
| Summary..... | 127 |
| To Learn More | 127 |
| 9 Model Building | 129 |
| Regression Models..... | 129 |
| Bivariate Dependence..... | 131 |
| Applying the Permutation Test..... | 131 |
| Models with a Single Predictor | 132 |

| | |
|---|-----|
| Comparing Two Regression Lines..... | 132 |
| Multipredictor Regression..... | 134 |
| Adaptive Regression..... | 136 |
| Applying the Bootstrap..... | 137 |
| Stata..... | 138 |
| Building a Model..... | 139 |
| Limitations of the Bootstrap..... | 140 |
| Prediction Error..... | 140 |
| Cross-Validation..... | 141 |
| Double Bootstrap..... | 141 |
| Validation..... | 141 |
| Metrics..... | 142 |
| Nearest Neighbors..... | 142 |
| Goodness of Fit..... | 143 |
| Using the Bootstrap for Model Validation..... | 144 |
| R Code..... | 145 |
| Cross-Validation..... | 145 |
| Summary..... | 146 |
| To Learn More..... | 146 |
| Exercises..... | 147 |

| | |
|--|-----|
| 10 Classification | 149 |
| Cluster Analysis..... | 149 |
| Classification..... | 151 |
| Decision Trees..... | 154 |
| Refining the Model..... | 155 |
| Decision Trees vs. Regression..... | 155 |
| Which Predictors?..... | 158 |
| Which Decision Tree Algorithm Is Best for Your Application?..... | 159 |
| Some Comparisons..... | 163 |
| Reducing the Rate of Misclassification..... | 163 |
| Boosting..... | 163 |
| AdaBoost Algorithm..... | 167 |
| Ensemble Methods..... | 167 |
| Comparison of Classification Tree Algorithms..... | 168 |
| Validation vs. Cross-Validation..... | 170 |
| Summary..... | 170 |
| To Learn More..... | 171 |
| Exercises..... | 172 |

| | |
|--|-----|
| 11 Restricted Permutations | 173 |
| Quasi Independence..... | 173 |
| Complete Factorials..... | 174 |

| | |
|--|-----|
| Synchronized Permutations..... | 175 |
| Generalizing These Results to Multiple Factors..... | 177 |
| Algorithms..... | 179 |
| Which Test Should We Use?..... | 180 |
| Model Validation..... | 180 |
| Exercises..... | 181 |
| <u>References</u> | 183 |
| <u>Appendix A: Basic Concepts in Statistics</u> | 201 |
| Additive vs. Multiplicative Models..... | 201 |
| Central Values..... | 201 |
| Combinations and Rearrangements..... | 202 |
| Dispersion..... | 202 |
| Frequency Distribution and Percentiles..... | 203 |
| Linear vs. Nonlinear Regression..... | 203 |
| Regression Methods..... | 204 |
| <u>Appendix B: Proof of Theorems</u> | 205 |
| <u>Subject Index</u> | 211 |

Preface

This text aspires to introduce statistical methodology to a wide audience, simply and intuitively, through resampling from the data at hand.

Practitioners, whether trained as statisticians or having the title *statistician* foisted upon them, soon realize the “normal” distribution almost never occurs in practice, there is no practical equivalent of standard error for measuring the precision of sample percentiles and variances, and it is virtually impossible to explain the meaning of regression coefficients to those whom you are supposed to be helping.

Readers of this text will learn to make use of distribution-free permutation tests, to estimate the precision of sample-based estimates using the bootstrap, and to replace arcane regression methods with the readily interpreted decision tree.

They’ll find the resampling methods—permutations, decision trees, and the bootstrap—easy to learn and easy to apply, for these methods require no mathematics beyond introductory high school algebra, yet are applicable to an exceptionally broad range of subject areas, as can be seen by the extensive list of cited applications in Chapter 1.

Practitioners and research workers in the biomedical, engineering, and social sciences, as well as advanced students in biology, business, dentistry, medicine, psychology, public health, and sociology, will find here a practical and easily grasped guide to descriptive statistics, estimation, and testing hypotheses.

Readers of my previous text for CRC, *Applying Statistics in the Courtroom*, will find the resampling methods invaluable at law, as they require the minimum of assumptions and lend themselves readily to interpretation by nonstatisticians. A hundred or more exercises included at the end of each chapter, plus dozens of thought-provoking questions, will serve the needs of both classroom and self-study. For those whose last statistics course was a long time ago, we’ve included a glossary of basic concepts in an appendix.

As statistics without software is as useless as Jell-O without a serving dish, we have provided R and Stata code for most resampling methods. We also have made available APL contributed by Valter Sundh, C++, Eviews, Gauss, Matlab contributed by Melanie Duffin, and SC code for many of the routines at our website: <http://statcourse.com/PGsoftware.htm>. A list of commercially available software for permutation tests and the bootstrap is provided in Chapter 3, and for decision trees in Section 11.6.