A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling

Phillip I. Good

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Contents

Preface ix

1 Wide Range of Applications 1
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
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
Original Sample 22
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 Comparing Two Populations 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 Multiple Variables 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 Experimental Design and Analysis 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 × 2s 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

Appendix A: Basic Concepts in Statistics 183
Additive vs. Multiplicative Models 183
Central Values 183
Combinations and Rearrangements 184
Dispersion 184
Frequency Distribution and Percentiles 185
Linear vs. Nonlinear Regression 185
Regression Methods 186

Appendix B: Proof of Theorems 187

References 193

Index 211