

Mathematical Statistics with Resampling and R

Laura Chihara
Carleton College

Tim Hesterberg
Google

Preface

Mathematical Statistics with Resampling and R is a one term undergraduate statistics textbook aimed at sophomores or juniors who have taken a course in probability (at the level of, for instance, Ross (2009), Ghahramani (2004) or Scheaffer and Young (2010)) but may not have had any previous exposure to statistics.

What sets this book apart from other mathematical statistics texts is the use of modern resampling techniques—permutation tests and bootstrapping. We begin with permutation tests and bootstrap methods before introducing classical inference methods. Resampling helps students understand the meaning of sampling distributions, sampling variability, P-values, hypothesis tests, and confidence intervals. We are inspired by the textbooks of Waldrop (1995) and Chance and Rossman (2005), two innovative introductory statistics books which also take a non-traditional approach in the sequencing of topics.

We believe the time is ripe for this book. Many faculty have learned resampling and simulation-based methods in graduate school and/or use them in their own work, and are eager to incorporate these ideas into a mathematical statistics course. Students and faculty today have access to computers that are powerful enough to perform resampling quickly.

A major topic of debate about the Mathematical Statistics course is how much theory to introduce. We want mathematically talented students to get excited about statistics, so we try to strike a balance between theory, computing and applications. We feel that it is important to demonstrate some rigor in developing some of the statistical ideas presented here, but that mathematical theory should not dominate the text. And of course, anytime additions are made to a syllabus, deletions must then also be made. Thus, some topics such as sufficiency, Fisher information and ANOVA have been omitted in order to make room for permutation testing, bootstrap and other modern computing methods (though we plan to make some of these omitted topics available as supplements on the text web page <https://sites.google.com/site/ChiharaHesterberg>).

We have compiled the definitions and theorems of the important probability distributions into an appendix (see Appendix B). Instructors who want to prove results on distributional theory can refer to that chapter. Instructors who wish to skip the theory can continue without interrupting the flow of the statistical discussion.

Incorporating resampling and bootstrapping methods requires that students use statistical software. We use R because it is freely available (<http://www.r-project.org/>), powerful, flexible, and a valuable tool in future careers. One of us works at Google where there is an explosion in the use of R, with more and more non-statisticians learning R (the statisticians already know it). We realize that the learning curve for R is high, but believe that the time invested in mastering R is worth the effort. We have written some basic materials on R that are available on the website for this text. We recommend that instructors work through the introductory worksheet with the students on the first or second day of the term, in a computer lab if possible. We also provide R script files with code found in the text, and additional examples.

Statistical computing is necessary in statistical practice and for people working with data in a wide variety of fields. There is an explosion of data—more and more data—and new computational methods are continuously being developed to handle this explosion. Statistics is an exciting field, dare we even say sexy?¹

Acknowledgements This textbook could not have been completed without the assistance of many colleagues and students. In particular, we would like to thank Professor Katherine St. Clair of Carleton College who bravely class-tested an early rough draft in her *Introduction to Statistical Inference* class during Winter 2010. In addition, Professor Julie Legler of St. Olaf College adopted the manuscript in her *Statistical Theory* class for Fall 2010. Both instructors and students provided valuable feedback that improved the exposition and content of this textbook.

We would also like to thank Siyuan (Ernest) Liu and Chen (Daisy) Sun, two Carleton College students, for solving many of the exercises and writing up the solutions with L^AT_EX.

Finally, the staff at Wiley, including Steve Quigley and Jackie Palmieri, provided valuable assistance in preparing the manuscript for press.

¹Try googling “statistics sexy profession.”

Contents

Preface	viii
1 Data and Case Studies	1
1.1 Case Study: Flight Delays	1
1.2 Case Study: Birth Weights of Babies	2
1.3 Case Study: Verizon Repair Times	3
1.4 Sampling	3
1.5 Parameters and Statistics	4
1.6 Case Study: General Social Survey	5
1.7 Sample Surveys	5
1.8 Case Study: Beer and Hot Wings	6
1.9 Case Study: Black Spruce Seedlings	7
1.10 Studies	7
1.11 Exercises	8
2 Exploratory Data Analysis	10
2.1 Basic Plots	10
2.2 Numeric Summaries	13
2.3 Boxplots	15
2.4 Quantiles and Normal Quantile Plots	16
2.5 Empirical Cumulative Distribution Functions	20
2.6 Scatter Plots	22
2.7 Skewness and kurtosis*	23
2.8 Exercises	24
3 Hypothesis Testing	28
3.1 Introduction to Hypothesis Testing	28
3.2 Hypotheses	29
3.3 Permutation Tests	31

3.3.1	Implementation issues	35
3.3.2	One-sided and two-sided tests	38
3.3.3	Other statistics	39
3.3.4	Assumptions	41
3.4	Contingency Tables	42
3.4.1	Permutation Testing for Independence	44
3.4.2	Chi-square reference distribution	47
3.5	Test of Independence	48
3.6	Test of homogeneity	51
3.7	Goodness-of-Fit: All Parameters Known	53
3.8	Goodness-of-Fit: Some Parameters Estimated	55
3.9	Exercises	57
4	Sampling Distributions	62
4.1	Sampling Distributions	62
4.2	Calculating Sampling Distributions	66
4.3	The Central Limit Theorem	68
4.3.1	CLT for Binomial Data	70
4.3.2	Continuity Correction for Discrete Random Variables	72
4.3.3	Accuracy of the Central Limit Theorem	73
4.3.4	CLT for Sampling Without Replacement	74
4.4	Exercises	74
5	The Bootstrap	80
5.1	Introduction to the Bootstrap	80
5.2	The Plug-in Principle	87
5.2.1	Estimating the population distribution	89
5.2.2	How useful is the bootstrap distribution?	90
5.3	Bootstrap Percentile Intervals	93
5.4	Two Sample Bootstrap	94
5.5	Other statistics	99
5.6	Bias	101
5.7	Monte Carlo Sampling - the “second bootstrap principle”	104
5.8	Accuracy of bootstrap distributions	104
5.9	How many bootstrap samples are needed?	108
5.10	Exercises	108

6	Estimation	112
6.1	Maximum Likelihood Estimation	112
6.1.1	Maximum Likelihood for Discrete distributions	112
6.1.2	Maximum Likelihood for Continuous Distributions	115
6.1.3	Maximum Likelihood for Multiple Parameters	118
6.2	Method of Moments	121
6.3	Properties of Estimators	123
6.3.1	Unbiasedness	123
6.3.2	Efficiency	126
6.3.3	Mean-square error	129
6.3.4	Consistency	130
6.3.5	Transformation Invariance	133
6.4	Exercises	134
7	Classical Inference: Confidence Intervals	138
7.1	Confidence Intervals for Means	138
7.1.1	Confidence Intervals for a Mean, σ Known	138
7.1.2	Confidence Intervals for a Mean, σ Unknown	142
7.1.3	Confidence Intervals for a Difference in Means	147
7.2	Confidence Intervals in General	152
7.2.1	Location and Scale Parameters*	155
7.3	One-Sided Confidence Intervals	157
7.4	Confidence Intervals for Proportions	159
7.4.1	The Agresti-Coull Interval for a Proportion	161
7.4.2	Confidence Interval for the Difference of Proportions	162
7.5	Bootstrap t Confidence Intervals	163
7.5.1	Comparing Bootstrap t and Formula t Confidence Intervals	167
7.6	Exercises	168
8	Classical Inference: Hypothesis Testing	175
8.1	Hypothesis Tests for Means and Proportions	175
8.1.1	One population	175
8.1.2	Comparing Two Populations	178
8.2	Type I and Type II errors	183
8.2.1	Type I errors	183
8.2.2	Type II errors and power	187
8.3	More on testing	191

8.3.1	On significance	191
8.3.2	Adjustments for multiple testing	192
8.3.3	P -values versus critical regions	193
8.4	Likelihood Ratio Tests	194
8.4.1	Simple Hypotheses and the Neyman-Pearson Lemma	194
8.4.2	Generalized Likelihood Ratio Tests	196
8.5	Exercises	198
9	Regression	203
9.1	Covariance	203
9.2	Correlation	206
9.3	Least Squares Regression	209
9.3.1	Regression Toward The Mean	212
9.3.2	Variation	213
9.3.3	Diagnostics	214
9.3.4	Multiple Regression	218
9.4	The Simple Linear Model	219
9.4.1	Inference for α and β	222
9.4.2	Inference for the Response	225
9.4.3	Comments about Assumptions for the Linear Model	228
9.5	Resampling Correlation and Regression	230
9.6	Logistic Regression	236
9.6.1	Inference for Logistic Regression	240
9.7	Exercises	243
10	Bayesian Methods	248
10.1	Bayes Theorem	248
10.2	Binomial Data, Discrete Prior Distributions	249
10.3	Binomial Data, Continuous Prior Distributions	255
10.4	Continuous Data	261
10.5	Sequential Data	263
10.6	Exercises	266
11	Additional Topics	271
11.1	Smoothed Bootstrap	271
11.1.1	Kernel Density Estimate	273
11.2	Parametric Bootstrap	276

11.3	The Delta Method	278
11.4	Stratified Sampling	281
11.5	Computational Issues in Bayesian Analysis	282
11.6	Monte Carlo Integration	283
11.7	Importance Sampling	286
11.7.1	Ratio Estimate for Importance Sampling	291
11.7.2	Importance Sampling in Bayesian Applications:	293
11.8	Exercises	297
Appendices		300
A	Review of Probability	300
A.1	Basic Probability	300
A.2	Mean and Variance	301
A.3	The Mean of a Sample of Random Variables	302
A.4	The Law of Averages	304
A.5	The Normal Distribution	304
A.6	Sums of Normal Random Variables	305
A.7	Higher moments and the moment generating function	306
B	Probability Distributions	308
B.1	The Bernoulli and Binomial Distributions	308
B.2	The Multinomial Distribution	308
B.3	The Geometric Distribution	310
B.4	The Negative Binomial Distribution	311
B.5	The Hypergeometric Distribution	312
B.6	The Poisson Distribution	313
B.7	The Uniform Distribution	314
B.8	The Exponential Distribution	314
B.9	The Gamma Distribution	315
B.10	The Chi-square Distribution	317
B.11	The Student's t Distribution	320
B.12	The Beta Distribution	321
B.13	The F Distribution	322
B.14	Exercises	324
C	Distributions Quick Reference	325

Solutions to Odd Exercises	328
Bibliography	334