Resampling Stats in MATLAB

Daniel T. Kaplan
Macalester College

Resampling Stats, Inc.
Arlington, Virginia
www.resample.com

To Maya, Tamar, Liat & Netta

ISBN 0-9672088-0-7

About the cover: *An image of a North Atlantic Grouper has been resampled many times, producing a population of fish. In the resampling, each block of the original image is replaced with a randomly selected block from elsewhere in the image, where the selected block is constrained to have a similar mean intensity and variance to the block being replaced. Cover Design: Tien Nguyen.*

Contents

Preface **vii**

Introduction to Statistical Inference 1

1 Sampling, Resampling, and Inference **1**
1.1 Sampling 3
1.2 Resampling 5
1.3 An Introduction to Probability 6

2 Making Statements with Precision: Confidence Intervals **19**

3 Testing Hypotheses with Data **35**
3.1 Concepts of Hypothesis Testing 35
3.2 Some Examples of Hypothesis Tests 39
3.2.1 Comparing two distributions 49
3.2.2 Independent resampling of two variables 56
3.2.3 Adjusting for multiple tests 57
3.2.4 Sample size and power 60

4 Updating our View: Bayesian Analysis **71**

5 Checking Resampling Results **79**
5.1 How many trials to use 79
5.2 How Much Data is Needed 81
5.3 Testing your programs 88

References **89**

Software Documentation **91**

A The Resampling Stats Commands **93**
Arithmetic . . . 93
Basic Descriptive Statistics . . . 95
between . . . 95
boxplt . . . 96
concat . . . 97
confintervals . . . 99
corr . . . 100
count . . . 100
dedup . . . 101
exclude . . . 102
expand . . . 103
exponential . . . 104
help . . . 105
histogram . . . 105
ismissing . . . 107
jab . . . 108
lambda . . . 109
length . . . 111
makerow . . . 112
max . . . 112
min . . . 113
mode . . . 113
multiples . . . 114
normal . . . 115
pause . . . 115
percentile . . . 116
plot . . . 117
proportion . . . 118
ranks . . . 119
recode . . . 120
regress . . . 121
resamp . . . 123
reverse . . . 124
round . . . 125
runs . . . 125
sample . . . 126
seed . . . 128
setmissing . . . 129

sepmatrix 130
shuffle 130
size 132
sort 132
starttally 133
std 133
tagsort 134
tally 135
twoway 136
uniform 137
urn 137
variance 138
weed 139
who 140

B Tutorial Introduction to MATLAB 141
Step 1: Starting MATLAB 141
Step 2: Defining Variables 142
Step 3: Using Variables 143
Step 4: Using Functions 144
Step 5: Making Vectors 145
Step 6: Arithmetic with Vectors 146
Step 7: Other Vector Operations 147
Step 8: Boolean Questions 147
Step 9: Loops and Repeating 148
Step 10: Conditional Expressions 149
Step 11: Saving Your Work 150
Step 12: Starting a New Session 150
Step 13: M-file Scripts 151
Step 14: M-file Functions 152
Step 15: Vectors and Matrices 154

C Reading and Saving Data 157
C.1 Importing External Data 157
C.2 Saving and Reading MATLAB Variables for Internal Use 162
C.3 Exporting MATLAB Results 163
C.4 Missing Data 163
C.5 Saving Figures 164

D Software installation & some technical matters **167**
D.1 Installing the Resampling Stats Software 167
D.1.1 Copying the Resampling Stats Files 167
D.1.2 Telling MATLAB where the files are 168
D.1.3 Using the MATLAB path browser 168
D.1.4 If someone else has installed Resampling Stats . . 169
D.1.5 Printing Numbers Nicely 170
D.1.6 Using the STARTRS script 170
D.2 Speed 171
D.2.1 Sample 171
D.2.2 Tallying results 172

Index **173**

List of Examples

1: Rolling a Die *8*
2: The Birthday Problem *12*
3: Annual Rainfall *15*
4: The Campaign Advisor States Confidently.... *20*
5: Uncertainty in the Mean *22*
6: Confidence in the extreme *24*
7: Housing prices: Confidence in the Median *25*
8: Confidence in the Distance *26*
9: Confidence in Correlations: Storks and Babies *27*
10: How Safe is the Space Shuttle? *31*
11: A Basketball Slump? *36*
12: Testing a Difference in Proportions *39*
13: Labor Trouble: Difference in Two Means *41*
14: Testing a difference in paired data *46*
15: Rolling a Die, revisited *49*
16: Grade Inflation *51*
17: Testing a correlation *56*
18: Discounting Multiple Comparisons *57*
19: Designing a Medical Study *61*
20: The Statistical Power of Polling *64*
21: The Basketball Slump Revisited *72*
22: Revisiting the Space Shuttle *74*
23: Enough light-bulb data? *85*

Preface

Resampling Stats is a system for carrying out computations in statistics and for conducting simulations. The computations relate to an area called "statistical inference" that deals with questions such as these:

- If an exit poll of 500 randomly selected voters in a national election shows that candidate A is favored by 41% of voters while candidate B trails with 35% of the vote, how confident can I be that candidate A will still be in the lead when all the votes are counted?
- A test of a blood-pressure reducing drug in 50 subjects shows that it reduces blood pressure by an average of 9.5 mmHg, whereas a placebo (a sugar pill) shows a reduction of 1.2 mmHg an a second group of 50. Am I justified in concluding that the drug is effective?
- If the space shuttle flies its first 24 flights without an accident do I have reason to believe that it is perfectly safe? If not, what is the accident rate I should use in planning future missions?
- An experiment in educational reform will give randomly selected families free tuition to private schools, while a control group of families will send their kids to public schools. The experiment is controversial and expensive; it's important to get meaningful results. How many families should be enrolled in the experiment?

Readers who have experience with statistics will recognize these questions as examples of the application of confidence intervals, hypothesis testing, and power computations. In conventional statistics courses students are taught how to answer questions like these using a certain theoretical apparatus (based on "Normal distribution theory", the t-distribution, and so on). If things go right in the course, students also learn how to interpret the answers to such questions and when there is not enough information to answer the posed questions. (For instance, in the second and fourth examples above there is not enough information.)

Resampling provides another, conceptually easier way to carry out the computations. In the theory of statistics, resampling is important

because it allows questions to be answered even in situations where the historically conventional methods do not apply. In the learning and teaching of statistics, resampling is valuable because it allows students to address the questions of statistical inference in a way where their intuition can be brought to bear, by designing and carrying out simple numerical experiments on the computer.

By making the computations more accessible, resampling has another important benefit: it allows students to move on to the important matters of how to interpret the numerical answers to their questions and how to know when there is not enough information to answer the question.

Resampling Stats was originally developed by Julian Simon during the period 1973-1990 as a stand-alone software package. As the benefits of the resampling approach to teaching statistics have become more apparent it seemed advisable to make the facilities of Resampling Stats available to a wider audience, and to allow users to employ Resampling Stats in a widely used computational environment.

There is a large community of people who use the MATLAB computer language. It is very widely used, for example, by engineering students and often used in teaching mathematics. MATLAB provides an integrated environment for technical computation: it provides facilities for drawing graphs, reading and saving data, and carrying out a tremendous range of numerical calculations. Since so many people already know MATLAB, or will need to learn it in order to carry out work in their chosen fields, MATLAB is a natural platform for Resampling Stats.

At the same time, we realize that for many students of Resampling Stats this will be their first encounter with MATLAB, and some will not use MATLAB for any other purpose. We have therefore worked hard to keep the original simplicity and ease of use of Resampling Stats. *We do not assume that you have any previous knowledge of* MATLAB. A tutorial in Appendix B can be used to get started for those who have no previous experience in MATLAB.

The body of this book is divided into two parts. First, there is an introduction to the issues and terms of statistical inference done mainly through examples. This introduction is thoroughly integrated with computer examples using Resampling Stats in MATLAB. In addition to showing how resampling can be used to answer the simple, standard statistical inference questions found in traditional introductory statistics textbooks, we show cases where traditional introductory methods do not apply but where resampling techniques are straightforward extensions of the simple cases. The examples introduce and cover both the "hypothesis testing" framework for statistical inference and the Bayesian approach.

The second part of the book is documentation for the various Resampling Stats functions in MATLAB. This is arranged as a reference rather than a tutorial. Appendices provide a tutorial introduction to MATLAB and show how to perform the important operation of reading data into the MATLAB program.

This book is intended mainly to introduce — using examples — the resampling methodology and the Resampling Stats in MATLAB software. We attempt to provide enough conceptual background and definition of statistical terms to make the book self contained. "Self contained" is not, however, the same thing as "systematic" or "comprehensive." This book does not cover all methods of analysis in statistical inference, nor does it do more than touch on the very important areas of experimental design, descriptive statistics, and exploratory data analysis. The treatment of these areas is largely independent of the mathematical methods — resampling vs. conventional formula — used for inference, although we believe that resampling is both more flexible and easier to learn.

Although the simple computer skills needed to use resampling are by no means trivial, we think they are far, far less formidable than the analytical mathematics that has been the bane of generations of statistics students who learned inference in the traditional way. Had computers been available 100 years ago, we think it likely that statistical inference would have developed with resampling as its foundation. As support for this entirely speculative statement, we note that one of the most important developments in traditional inference theory, the t-distribution, was developed at the turn of the last century by William Gosset based on resampling techniques (and tedious labor on hand calculators).

Whatever their virtues, the modern computer-intensive techniques and the Resampling Stats software do not automatically translate data into answers. Instead, they allow you to design and carry out computer experiments to find answers to your own questions. The examples in this book show how the software is used and illustrate some common types of experiments, but they do not cover the most important cases: those that specifically address the questions you want to ask about your data. We hope that Resampling Stats will give you the facility to answer these important questions about your own data.

We would like to thank Peter Bruce, Dan Hornbach, Paul Alper and Rob Leduc for their help in the writing of this book.

St. Paul, Minnesota, July 1999