INTRODUCTION TO STATISTICS THROUGH RESAMPLING METHODS AND MICROSOFT OFFICE EXCEL®

Phillip I. Good
INTRODUCTION TO STATISTICS THROUGH RESAMPLING METHODS AND MICROSOFT OFFICE EXCEL®
Contents

Preface xi

1. Variation (or What Statistics Is All About) 1
   1.1. Variation 1
   1.2. Collecting Data 2
   1.3. Summarizing Your Data 3
       1.3.1 Learning to Use Excel 4
   1.4. Reporting Your Results: the Classroom Data 7
       1.4.1 Picturing Data 10
       1.4.2 Displaying Multiple Variables 10
       1.4.3 Percentiles of the Distribution 15
   1.5. Types of Data 20
       1.5.1 Depicting Categorical Data 21
       1.5.2 From Observations to Questions 23
   1.6. Measures of Location 23
       1.6.1 Which Measure of Location? 25
       1.6.2 The Bootstrap 27
   1.7. Samples and Populations 30
       1.7.1 Drawing a Random Sample 32
       1.7.2 Ensuring the Sample is Representative 34
   1.8. Variation—Within and Between 34
   1.9. Summary and Review 36

2. Probability 39
   2.1. Probability 39
       2.1.1 Events and Outcomes 41
       2.1.2 Venn Diagrams 41
   2.2. Binomial 43
       2.2.1 Permutations and Rearrangements 45
       2.2.2 Back to the Binomial 47
4.2.3 Permutation Monte Carlo 95
4.2.4 Two-Sample \( t \)-Test 97

4.3. Which Test Should We Use? 97
4.3.1 \( p \) Values and Significance Levels 98
4.3.2 Test Assumptions 98
4.3.3 Robustness 99
4.3.4 Power of a Test Procedure 100
4.3.5 Testing for Correlation 101

4.4. Summary and Review 104

5. Designing an Experiment or Survey 105
5.1. The Hawthorne Effect 106
5.1.1 Crafting an Experiment 106
5.2. Designing an Experiment or Survey 108
5.2.1 Objectives 109
5.2.2 Sample from the Right Population 110
5.2.3 Coping with Variation 112
5.2.4 Matched Pairs 113
5.2.5 The Experimental Unit 114
5.2.6 Formulate Your Hypotheses 114
5.2.7 What Are You Going to Measure? 115
5.2.8 Random Representative Samples 116
5.2.9 Treatment Allocation 117
5.2.10 Choosing a Random Sample 118
5.2.11 Ensuring that Your Observations are Independent 119

5.3. How Large a Sample? 120
5.3.1 Samples of Fixed Size 121
- Known Distribution 122
- Almost Normal Data 125
- Bootstrap 127
5.3.2 Sequential Sampling 129
- Stein’s Two-Stage Sampling Procedure 129
- Wald Sequential Sampling 129
- Adaptive Sampling 133

5.4. Meta-Analysis 134
5.5. Summary and Review 135

6. Analyzing Complex Experiments 137
6.1. Changes Measured in Percentages 137
6.2. Comparing More Than Two Samples 138
6.2.1 Programming the Multisample Comparison with Excel 139
6.2.2 What Is the Alternative? 141
6.2.3 Testing for a Dose Response or Other Ordered Alternative 141
6.3. Equalizing Variances 145
6.4. Stratified Samples 147
6.5. Categorical Data 148
   6.5.1 One-Sided Fisher’s Exact Test 150
   6.5.2 The Two-Sided Test 151
   6.5.3 Multinomial Tables 152
   6.5.4 Ordered Categories 153
6.6. Summary and Review 154

7. Developing Models 155
   7.1. Models 155
      7.1.1 Why Build Models? 156
      7.1.2 Caveats 158
   7.2. Regression 159
      7.2.1 Linear Regression 160
   7.3. Fitting a Regression Equation 161
      7.3.1 Ordinary Least Squares 162
         • Types of Data 166
      7.3.2 Least Absolute Deviation Regression 168
      7.3.3 Errors-in-Variables Regression 168
      7.3.4 Assumptions 171
   7.4. Problems with Regression 172
      7.4.1 Goodness of fit versus prediction 172
      7.4.2 Which Model? 173
      7.4.3 Measures of Predictive Success 174
      7.4.4 Multivariable Regression 175
   7.5. Quantile Regression 182
   7.6. Validation 183
      7.6.1 Independent Verification 183
      7.6.2 Splitting the Sample 184
      7.6.3 Cross-Validation with the Bootstrap 185
   7.7. Classification and Regression Trees 186
   7.8. Data Mining 190
   7.9. Summary and Review 193
8. Reporting Your Findings 195
  8.1. What to Report 195
  8.2. Text, Table, or Graph? 199
  8.3. Summarizing Your Results 200
    8.3.1 Center of the Distribution 201
    8.3.2 Dispersion 203
  8.4. Reporting Analysis Results 204
    8.4.1 p Values? Or Confidence Intervals? 205
  8.5. Exceptions Are the Real Story 206
    8.5.1 Nonresponders 206
    8.5.2 The Missing Holes 207
    8.5.3 Missing Data 207
    8.5.4 Recognize and Report Biases 208
  8.6. Summary and Review 209

9. Problem Solving 211
  9.1. The Problems 211
  9.2. Solving Practical Problems 215
    9.2.1 The Data’s Provenance 215
    9.2.2 Inspect the Data 216
    9.2.3 Validate the Data Collection Methods 217
    9.2.4 Formulate Hypotheses 217
    9.2.5 Choosing a Statistical Methodology 218
    9.2.6 Be Aware of What You Don’t Know 218
    9.2.7 Qualify Your Conclusions 218

Appendix: An Microsoft Office Excel Primer 221

Index to Excel and Excel Add-In Functions 227

Subject Index 229
INTENDED FOR CLASS USE OR SELF-STUDY, this text aspires to introduce statistical methodology to a wide audience, simply and intuitively, through resampling from the data at hand.

The resampling methods—permutations and the bootstrap—are easy to learn and easy to apply. They require no mathematics beyond introductory high-school algebra, yet are applicable in an exceptionally broad range of subject areas.

Introduced in the 1930s, the numerous, albeit straightforward calculations resampling methods require were beyond the capabilities of the primitive calculators then in use. They were soon displaced by less powerful, less accurate approximations that made use of tables. Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role and table lookup is an anachronism.

Physicians and physicians in training, nurses and nursing students, business persons, business majors, research workers, and students in the biological and social sciences will find here a practical and easily grasped guide to descriptive statistics, estimation, testing hypotheses, and model building.

For advanced students in biology, dentistry, medicine, psychology, sociology, and public health, this text can provide a first course in statistics and quantitative reasoning.

For mathematics majors, this text will form the first course in statistics, to be followed by a second course devoted to distribution theory and asymptotic results.

Hopefully, all readers will find my objectives are the same as theirs: To use quantitative methods to characterize, review, report on, test, estimate, and classify findings.

Warning to the autodidact: You can master the material in this text without the aid of an instructor. But you may not be able to grasp even
the more elementary concepts without completing the exercises. Whenever and wherever you encounter an exercise in the text, stop your reading and complete the exercise before going further.

You’ll need to download and install several add-ins for Excel to do the exercises, including BoxSampler, Ctree, DDXL, Resampling Statistics for Excel, and XLStat. All are available in no-charge trial versions. Complete instructions for doing the installations are provided in Chapter 1. For those brand new to Excel itself, a primer is included as an Appendix to the text.

For a one-quarter short course, I’d recommend taking students through Chapters 1 and 2 and part of Chapter 3. Chapters 3 and 4 would be completed in the winter quarter along with the start of chapter 5, finishing the year with Chapters 5, 6, and 7. Chapters 8 and 9 on “Reporting Your Findings” and “Problem Solving” convert the text into an invaluable professional resource.

An Instructor’s Manual is available to qualified instructors and may be obtained by contacting the Publisher. Please visit ftp://ftp.wiley.com/public/sci_tech_med/introduction_statistics/ for instructions on how to request a copy of the manual.

Twenty-eight or more exercises included in each chapter plus dozens of thought-provoking questions in Chapter 9 will serve the needs of both classroom and self-study. The discovery method is utilized as often as possible, and the student and conscientious reader are forced to think their way to a solution rather than being able to copy the answer or apply a formula straight out of the text. To reduce the scutwork to a minimum, the data sets for the exercises may be downloaded from ftp://ftp.wiley.com/public/sci_tech_med/statistics_resampling.

If you find this text an easy read, then your gratitude should go to Cliff Lunneborg for his many corrections and clarifications. I am deeply indebted to the students in the Introductory Statistics and Resampling Methods courses that I offer on-line each quarter through the auspices of statistics.com for their comments and corrections.

Phillip I. Good
Huntington Beach, CA
frere_until@hotmail.com
If there were no variation, if every observation were predictable, a mere repetition of what had gone before, there would be no need for statistics.

1.1. VARIATION
We find physics extremely satisfying. In high school, we learned the formula \( S = VT \), which in symbols relates the distance traveled by an object to its velocity multiplied by the time spent in traveling. If the speedometer says 60 miles an hour, then in half an hour you are certain to travel exactly 30 miles. Except that during our morning commute, the speed we travel is seldom constant.

In college, we had Boyle’s law, \( V = KT/P \), with its tidy relationship between the volume \( V \), temperature \( T \), and pressure \( P \) of a perfect gas. This is just one example of the perfection encountered there. The problem was we could never quite duplicate this (or any other) law in the freshman physics laboratory. Maybe it was the measuring instruments, our lack of familiarity with the equipment, or simple measurement error, but we kept getting different values for the constant \( K \).

By now, we know that variation is the norm. Instead of getting a fixed, reproducible \( V \) to correspond to a specific \( T \) and \( P \), one ends up with a distribution of values instead as a result of errors in measurement. But we also know that with a large enough sample, the mean and shape of this distribution are reproducible.

That’s the good news: Make astronomical, physical, or chemical measurements and the only variation appears to be due to observational error. But try working with people.

Anyone who has spent any time in a schoolroom, whether as a parent or as a child, has become aware of the vast differences among individuals.
Our most distinct memories are of how large the girls were in the third grade (ever been beat up by a girl?) and the trepidation we felt on the playground whenever teams were chosen (not right field again!). Much later, in our college days, we were to discover there were many individuals capable of devouring larger quantities of alcohol than we could without noticeable effect, and a few, mostly of other nationalities, whom we could drink under the table.

Whether or not you imbibe, we’re sure you’ve had the opportunity to observe the effects of alcohol on others. Some individuals take a single drink and their nose turns red. Others can’t seem to take just one drink.

The majority of effort in experimental design, the focus of Chapter 5 of this text, is devoted to finding ways in which this variation from individual to individual won’t swamp or mask the variation that results from differences in treatment or approach. It’s probably safe to say that what distinguishes statistics from all other branches of applied mathematics is that it is devoted to characterizing and then accounting for variation.

### SOURCES OF VARIATION

You catch three fish. You heft each one and estimate its weight; you weigh each one on a pan scale when you get back to dock, and you take them to a chemistry laboratory and weigh them there. Your two friends on the boat do exactly the same thing. (All but Mike; the chem professor catches him and calls campus security. This is known as missing data.)

The 26 weights you’ve recorded (3 \( \times \) 3 \( \times \) 3 – 1 when they nabbed Mike) differ as result of measurement error, observer error, differences among observers, differences among measuring devices, and differences among fish.

### 1.2. COLLECTING DATA

The best way to observe variation is for you, the reader, to collect some data. But before we make some suggestions, a few words of caution are in order: 80% of the effort in any study goes into data collection and preparation for data collection. Any effort you don’t expend goes into cleaning up the resulting mess.

We constantly receive letters and E-mails asking which statistic we would use to rescue a misdirected study. There is no magic formula, no secret procedure known only to PhD statisticians. The operative phrase is GIGO: Garbage In, Garbage Out. So think carefully before you embark on your collection effort. Make a list of possible sources of variation and see whether you can eliminate any that are unrelated to the objectives of
your study. If midway through, you think of a better method—don’t use it. Any inconsistency in your procedure will only add to the undesired variation.

Let’s get started. Here are three suggestions. Before continuing with your reading, follow through on at least one of them or an equivalent idea of your own, as we will be using the data you collect in the very next section:

1. Measure the height, circumference, and weight of a dozen humans (or dogs, or hamsters, or frogs, or crickets).

2. Time some tasks. Record the times of 5–10 individuals over three track lengths (say 50 meters, 100 meters, and a quarter mile). Because the participants (or trial subjects) are sure to complain they could have done much better if only given the opportunity, record at least two times for each study subject. (Feel free to use frogs, hamsters, or turtles in place of humans as runners to be timed. Or to replace foot races with knot tying, bandaging, or putting on or taking off a uniform.)

3. Take a survey. Include at least three questions and survey at least 10 subjects. All your questions should take the form “Do you prefer A to B? Strongly prefer A, slightly prefer A, indifferent, slightly prefer B, strongly prefer B.” For example, “Do you prefer Britney Spears to Jennifer Lopez?” or “Would you prefer spending money on new classrooms rather than guns?”

### SOURCES OF VARIATION

- Characteristics of the observer(s)
- Characteristics of the environment in which observations are made
- Characteristics of the measuring device(s)
- Characteristics of the subjects or objects observed

**Exercise 1.1.** Collect data as described above. Before you begin, write down a complete description of exactly what you intend to measure and how you plan to make your measurements. Make a list of all potential sources of variation. When your study is complete, describe what deviations you had to make from your plan and what additional sources of variation you encountered.

### 1.3. SUMMARIZING YOUR DATA

Learning how to adequately summarize one’s data can be a major challenge. Can it be explained with a single number like the median? The
median is the middle value of the observations you have taken, so that half of the data have a smaller value and half have a greater value. Take the observations 1.2, 2.3, 4.0, 3, and 5.1. The observation 3 is the one in the middle. If we have an even number of observations such as 1.2, 2.3, 3, 3.8, 4.0, and 5.1, then the best one can say is that the median or midpoint is a number (any number) between 3 and 3.8. Now, a question for you: What are the median values of the measurements you made?

Hopefully, you’ve already collected data as described in Section 1.2; otherwise, face it, you are behind. Get out the tape measure and the scales. If you conducted time trials, use those data instead. Treat the observations for each of the three distances separately.

If you conducted a survey, we have a bit of a problem. How does one translate “I would prefer spending money on new classrooms rather than guns” into a number a computer can add and subtract? There is more one way to do this, as we’ll discuss in what follows under the heading, “Types of Data.” For the moment, assign the number 1 to “Strongly prefer classrooms,” the number 2 to “Slightly prefer classrooms,” and so on.

1.3.1. Learning to Use Excel

Calculating the value of a statistic is easy enough when we’ve only 1 or 2 observations, but a major pain when we have 10 or more. And as for drawing graphs—one of the best ways to summarize your data—we’re no artists. Let the computer do the work.

We’re going to need the help of Excel, a spreadsheet program with many built-in statistics and graphics functions. We’ll assume that you already have Microsoft Office Excel installed and have some familiarity with its use.\(^1\) To enter the observations 1.2, 2.3, 4.0, 3, and 5.1, simply type these values down the first column starting in the third row. Notice in Fig. 1.1 that we’ve put a description of the column in the second row. The first row is reserved for a more lengthy description of the project should one be required.

In Fig. 1.1, we’ve begun in Row 8 to start the computation of the median of our data. Here are the steps we went through:

1. Type the first data element (1.2 in this example) in the third row of the first column.
2. Press the “Enter” key to go to the next row.

\(^1\) If you’re an absolute beginner, we’ve included an Appendix to the text to help you get started. If you already own and are familiar with some other statistics package or spreadsheet, feel free to use it instead. The objective of this text is to help you understand and make use of basic statistics principles. Excel is merely a convenient tool.
3. Repeat steps 1 and 2 until all the data are entered.
4. Use your mouse to depress the button in the row.

5. Depress the down arrow next to the word SUM and select “More Functions” from the resultant display (Fig. 1.2).
6. Select “Statistical” from the Function category menu and “Median” from the Function name menu.
7. Press “OK” or the “Enter” key to learn that the median of the five numbers we entered is 2.65.

The median of a sample tells us where the center of a set of observations is, but it provides no information about the variability of our observations, and variation is what statistics is all about. Pictures tell the story the best.

In Section 1.4, we’ll consider some data on heights I collected while teaching sixth-graders mathematics. The one-way strip chart or dotplot (Fig. 1.3) created with the aid of Data Desk/XL², an Excel add-in, reveals that the minimum of this particular set of data is approximately 137 cm

² A trial version may be downloaded from http://www.datadesk.com/products/data_analysis/ddxl/.
and the maximum approximately 167 cm. Each dot in this strip chart corresponds to an observation. Blotches correspond to multiple observations. The range over which these observations extend is 167–137, or 30.

By the way, DataDesk/XL is just one of a hundred or more programs that can add in capabilities to Excel. We’ll be using several such add-ins to carry out the necessary calculations to complete this course.
A weakness of Fig. 1.3 is that it’s hard to tell exactly what the values of the various percentiles are. A glance at the box and whiskers plot (Fig. 1.4) created with the aid of XLStat (Addinsoft, 2004), a second Excel add-in, tells us that the median of the classroom data described in Section 1.4 is 153.5 cm, the mean is 151.6 cm, and the interquartile range (the “box”) is close to 14 cm. The minimum and maximum of the sample are located at the ends of the “whiskers.”

In Section 1.4, you’ll learn how to create these and other graphs.

1.4. REPORTING YOUR RESULTS: THE CLASSROOM DATA
Imagine you are in the sixth grade and you have just completed measuring the heights of all your classmates.

Once the pandemonium has subsided, your instructor asks you and your team to prepare a report summarizing your results.

Actually, you have two sets of results. The first set consists of the measurements you made of you and your team members, reported in centimeters, 148.5, 150.0, and 153.0. (Kelly is the shortest, incidentally, and you are the tallest.) The instructor asks you to report the minimum, the

---

3 A trial version may be downloaded from http://www.xlstat.com/download.htm.
median, and the maximum height in your group. This part is easy, or at least it’s easy once you look the terms up in the glossary of your textbook and discover that minimum means smallest, maximum means largest, and median is the one in the middle. Conscientiously, you write these definitions down—they could be on a test.

In your group, the minimum height is 148.5 centimeters, the median is 150.0 centimeters, and the maximum is 153.0 centimeters.

Your second assignment is more challenging. The results from all your classmates have been written on the blackboard—all 22 of them.

141, 156.5, 162, 159, 157, 143.5, 154, 158, 140, 142, 150, 148.5, 138.5, 161, 153, 145, 147, 158.5, 160.5, 167.5, 155, 137

You copy the figures neatly into the first column of an Excel worksheet as described in the previous section. Next, you brainstorm with your teammates. Nothing. Then John speaks up—he’s always interrupting in class. Shouldn’t we put the heights in order from smallest to largest? “Of course,” says the teacher, “you should always begin by ordering your observations.”

You go to the Excel menu bar as shown in Fig. 1.5 and access the “sort” command from the “data” menu. As a result, your data are now in sorted in order from smallest to largest:

![Microsoft Excel - Book1](image)

**FIGURE 1.5** Accessing the sort command.
“I know what the minimum is,” you say—come to think of it, you are always blurting out in class, too, “137 millimeters, that’s Tony.”

“The maximum, 167.5, that’s Pedro, he’s tall,” hollers someone from the back of the room.

As for the median height, the one in the middle is just 153 centimeters (or is it 154)? What does Excel tell us? As illustrated in Fig. 1.6, we need to do the following to find out:

1. Put our cursor in the first empty cell after the data; A25 in our example.
2. Click the = key on the formula menu bar.
3. Select “median” by using the down arrow ▼ on the formula bar.

**FIGURE 1.6** Computing the median of the classroom data.
4. Use the cursor to select the data range or enter the data range using the form shown in Fig. 1.6 as A3:A24.

5. Press OK.

The result 153.5 will appear in cell A25.

Actually, the median could be any number between 153 and 154, but it is a custom among statisticians, honored by Excel, to report the median as the value midway between the two middle values, when the number of observations is even.

1.4.1. Picturing Data

The preceding scenario was a real one. The results reported here, especially the pandemonium, were obtained by my sixth grade homeroom at St. John’s Episcopal School in Rancho Santa Marguarite, CA. The problem of a metric tape measure was solved by building their own from string and a meter stick.

My students at St. John’s weren’t through with their assignments. It was important for them to build on and review what they’d learned in the fifth grade, so I had them draw pictures of their data. Not only is drawing a picture fun, but pictures and graphs are an essential first step toward recognizing patterns.

We begin by downloading a trial copy of DataDesk/XL from the website http://www.datadesk.com/products/data_analysis/downloads/ddxl.cfm. Note the folder to which you downloaded the program.

To install this add-in, pull down the Excel Tools menu, select “add-ins,” and then browse the various folders on the hard disk until you locate the DDXL add-in. Once DDXL is added, a new pull-down menu, labeled DDXL will appear on the menu bar as shown in Fig. 1.7.

After selecting “Charts and Plots” as depicted in Fig. 1.7, we complete the Charts and Plots Dialog shown in Fig. 1.8. Note that among the other possible headings under “Function type” are Box Plot and Histogram.

We click “OK”, and Fig. 1.9 reveals the end result. As a by-product, the numeric values of various sample statistics are displayed as well as the dotplot.

Exercise 1.2. Generate a dot plot and a box plot for one of the data sets you gathered in your initial assignment. Write down the values of the median, minimum, and maximum that you can infer from the box plot.

1.4.2. Displaying Multiple Variables

I’d read, but didn’t quite believe, that one’s arm span is almost exactly the same as one’s height. To test this hypothesis, I had my sixth graders get
CHAPTER 1 VARIATION (OR WHAT STATISTICS IS ALL ABOUT)

FIGURE 1.7 Selecting charts and plots from the DDXL menu.

FIGURE 1.8 Selecting the type of graph desired.