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**

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

**2. Probability**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Probability</td>
<td>39</td>
</tr>
<tr>
<td>2.1.1 Events and Outcomes</td>
<td>41</td>
</tr>
<tr>
<td>2.1.2 Venn Diagrams</td>
<td>41</td>
</tr>
<tr>
<td>2.2. Binomial</td>
<td>43</td>
</tr>
<tr>
<td>2.2.1 Permutations and Rearrangements</td>
<td>45</td>
</tr>
<tr>
<td>2.2.2 Back to the Binomial</td>
<td>47</td>
</tr>
</tbody>
</table>
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
  8.1. What to Report
  8.2. Text, Table, or Graph?
  8.3. Summarizing Your Results
    8.3.1 Center of the Distribution
    8.3.2 Dispersion
  8.4. Reporting Analysis Results
    8.4.1 \( p \) Values? Or Confidence Intervals?
  8.5. Exceptions Are the Real Story
    8.5.1 Nonresponders
    8.5.2 The Missing Holes
    8.5.3 Missing Data
    8.5.4 Recognize and Report Biases
  8.6. Summary and Review

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

Appendix: An Microsoft Office Excel Primer

Index to Excel and Excel Add-In Functions

Subject Index
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.