# Overview of

# <u>Statistics: UnLocking the Power of Data</u>


## By Lock, Lock, Lock, Lock, and Lock



**To be published by John Wiley and Sons, Inc.**
**Available for class testing.**


Sir R.A. Fisher said of simulation and permutation methods in 1936:

> "Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

These methods, too 'tedious' to apply in 1936, are now readily accessible. As George Cobb (2007) wrote in his lead article for the journal *Technology Innovations in Statistical Education*,

> "... despite broad acceptance and rapid growth in enrollments, the consensus curriculum is still an unwitting prisoner of history. What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs. This machinery was once necessary, because the conceptually simpler alternative based on permutations was computationally beyond our reach. Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course."

It is our hope that the textbook we are writing will help move the introductory statistics curriculum in the directions advocated by Professor Cobb. We use ideas such as randomization tests and bootstrap intervals to introduce the fundamental ideas of statistical inference. These methods are surprisingly intuitive to novice students and, with proper use of computer support, are accessible at very early stages of a course. Our text introduces statistical inference through these resampling methods, not only because these methods are becoming increasingly important for statisticians in their own right but also because randomization methods are outstanding in building students' conceptual understanding of the key ideas.

Our text includes the more traditional methods such as t-tests, chi-square tests, etc., but only after students have developed a strong intuitive understanding of inference through randomization methods. At this point students have a conceptual understanding and appreciation for the results they can then compute using the more traditional methods. We believe that this approach helps students realize that although the formulae may take different forms for different types of data, the conceptual framework underlying most statistical methods remains the same.

Furthermore, our experience has been that after using these new methods in intuitive ways to introduce the core ideas, students understand and can move quickly through most of the standard techniques. Our goal is a text that gently moves the curriculum in innovative ways while still looking relatively familiar. Instructors won't need to completely abandon their current syllabi and students will be well-prepared for more traditional follow-up courses.

The text is designed for use in a one-semester introductory statistics course taught at the 100-level. The focus throughout is on data analysis and the primary goal is to enable students to effectively collect data, analyze data, and interpret conclusions drawn from data. The text is driven by real data and real applications. Although the only prerequisite is a minimal working knowledge of algebra, students completing the course should be able to accurately interpret statistical results and to analyze straightforward data sets. The text is designed to give students a sense of the power of data analysis; our hope is that many students learning from this book will want to continue developing their statistical knowledge.

Students who learn from this text should finish with

- A solid conceptual understanding of the key concepts of statistical inference: estimation with intervals and testing for significance.
- The ability to do straightforward data analysis, using either traditional methods or modern resampling methods.
- Experience using technology to perform a variety of different statistical procedures.
- An understanding of the importance of data collection, the ability to recognize limitations in data collection methods, and an awareness of the role that data collection plays in determining the scope of inference.
- The knowledge of which statistical methods to use in which situations and the ability to interpret the results effectively and in context.
- An awareness of the power of data analysis.

<u>Technology</u>
A set of online interactive dynamic tools, called ***StatKey***, is fully integrated with the text and is available on the web to illustrate key ideas and provide support for computer intensive procedures. Other than this optional user-friendly set of web tools, the text is not tied to any specific statistical software package. Output from a variety of different packages is regularly displayed so that students become comfortable reading output in different forms. Companion manuals (both print and online) are available to provide specific computing guidance for common statistical packages. The text uses many real datasets and all data is electronically available in multiple formats.

<u>Examples and Exercises</u>
Applications in the text are drawn from a wide variety of disciplines. There are enough problems and examples so that individual instructors can focus on specific areas if desired. Topics in examples and problems have been chosen primarily on the basis of perceived interest to students and instructors. Problems and exercises are plentiful and span a very wide range of difficulty levels, from very straightforward short answer problems to extended projects.

<u>Essential Synthesis</u>
One of the drawbacks of many current texts is the fragmentation of ideas into disjoint pieces. While the segmentation helps students understand the individual pieces, we believe that integration of the parts into a coherent whole is also essential. To address this we have sections called *Essential Synthesis* at the end of each unit, in which students are asked to take a step back and look at the big picture. We hope that these integration sections will help to prepare students for the kind of statistical thinking they will most likely encounter after finishing the course.

<u>What the text is not</u>

- The full emphasis of the text is on developing the skills students need to effectively describe and analyze data. In particular, coverage of formal probability theory is minimal and appears at the end of the text.
- The text emphasizes concepts and does not take a "cookbook" approach. Techniques and skills will be developed, but not at the expense of understanding the ideas.
- The text is not purely conceptual. It is not just a "statistical literacy" book. It is designed to enable students to accurately describe and analyze data.
- The text deals seriously with fundamental ideas of statistics, but does not assume calculus or higher level mathematical skills. It is a user's guide to statistics, appropriate for any course in which the goal is to help students learn how to use statistics.

**Overview of the Table of Contents**

Unit A: Data includes chapters on data production (Chapter 1) and data description (Chapter 2). The first chapter has a dual role of introducing standard terminology about sampling, observational studies, and experiments, along with providing motivation for how statistics can be used to address real questions. Having this chapter early allows us to provide a "big picture" with compelling examples and continually return to questions about how data are collected and how that affects the subsequent analysis, interpretations, and scope of conclusions. Chapter 2 is about summarizing data and is relatively traditional. By the end of that chapter students should be able to summarize categorical and quantitative variables numerically and graphically, both for single variables and relationships between two variables.

Unit B: Understanding Inference introduces students to the key ideas of statistical inference. This is where most of the new computer-intensive approaches appear. Chapter 3 introduces confidence intervals based on bootstrap samples. Chapter 4 introduces significance tests using randomization distributions. By the end of these two chapters, students should be able to interpret a confidence interval, to state and understand hypotheses for a test, and to understand and interpret a p-value from a test. Our approach here is fairly general, so that students see these methods applied to many different parameters and data situations. The easy-to-use online applets in ***StatKey*** provide support throughout the course, but will be particularly helpful in this unit.

Unit C: Inference for Means and Proportions covers techniques for computing confidence intervals and hypothesis tests for means and proportions based on normal and t-distributions. The first chapter, Chapter 5, introduces the normal distribution as a model to summarize the distributions that students have been working with in Unit B. Students see how to use calculations based on the normal distribution to obtain a confidence interval or p-value in a general setting. Chapter 6 covers the traditional "shortcut" methods for doing tests and intervals for means and proportions in one and two sample settings. From talking with many instructors we understand that there is no consensus on an ideal ordering for doing these topics. We have designed Chapter 6 to have many short sections that allow instructors to easily adapt to any preferred order. Since students have already grappled with the fundamental ideas of intervals and tests in Unit B, instructors should be able to move through the sections of Chapter 6 rather quickly and in whatever order suits their own tastes.

Unit D: Inference for Multiple Parameters includes somewhat more advanced methods that typically involve more than one or two parameters. The first three chapters are designed to be covered in any order. Chapter 7 looks at chi-square tests for a single categorical variable (goodness-of-fit) or a relationship between two categorical variables (contingency table). Chapter 8 extends inference for two means to look at ANOVA to compare several means. Chapter 9 does a more formal treatment of inference for regression models based on a single predictor. Chapter 10 extends these ideas to give a brief look at regression models with multiple predictors.

The text ends with a final section called Essential Synthesis: The Big Picture in which students are asked to put all the pieces together. This includes case studies illustrating the entire process of using statistics to answer a research question; asking students to indicate the appropriate method of data collection, analysis technique, or conclusion to be drawn in specific situations; and other examples and problems synthesizing and applying what has been learned throughout the text.

The topics described above are covered in 10 chapters with a total of 46 sections, which we believe will work well for a typical one semester course. Depending on level, instructors may choose to combine some of the shorter sections (especially in Chapter 6) or skip some of the more advanced chapters/sections in Unit D.

At the very end of the text is a Chapter 11 on Probability, which is written to be flexible enough to be inserted when and if instructors wish.

**About the Authors**:

The five authors of the text all bring different strengths to this project. In addition, they think it is pretty cool that they are all in the same immediate family. In addition to the interests listed below, all authors share a deep interest in statistics education.

Robin H. Lock is Burry Professor of Statistics in the Department of Mathematics, Computer Science, and Statistics at St. Lawrence University. He is a Fellow of the American Statistical Association, past Chair of the Joint MAA-ASA Committee on Teaching Statistics, a member of the committee that developed GAISE (Guidelines for Assessment and Instruction in Statistics Education), and a member of the Consortium for the Advancement of Undergraduate Statistics Education, CAUSE. He has won the national Mu Sigma Rho Statistics Education award and numerous awards for presentations on statistics education at national conferences. He brings to the project an insider's understanding of national trends in statistics education.

Patti Frazer Lock is Cummings Professor of Mathematics in the Department of Mathematics, Computer Science and Statistics at St. Lawrence University. She is a member of the Committee on the Undergraduate Program in Mathematics of the Mathematics Association of American, is on the Editorial Board of PRIMUS Journal, and is a consultant for Project NExT. She is a co-author of several Wiley mathematics texts. As a mathematician who has taught many sections of introductory statistics, she brings to the project an awareness of what resources a mathematician needs in order to teach an effective statistics course.

Kari Lock Morgan completed her Ph. D. in the Department of Statistics at Harvard University in May 2011 and started a faculty position at Duke University in Fall 2011. She has taught a variety of statistics classes, including a special course for graduate students on "The Art and Practice of Teaching Statistics", and has co-developed a new 100-level course at Harvard designed to make statistics enjoyable and applicable to real life. At Harvard, she won the Derek C. Bok Award for Excellence in the Teaching of Undergraduates. She has a particular interest in causal inference, statistics education, and applications of statistics in psychology, education, and health.

Eric F. Lock is in his fifth year as a Ph.D. candidate in the Department of Statistics and Operations Research at the University of North Carolina, and expects to complete his Ph.D. in May 2012. He has been an instructor and instructional assistant for multiple introductory statistics courses, ranging from very traditional to more progressive, and has won a teaching award at UNC. He has a particular interest in machine learning and the analysis of high-dimensional data, and has conducted research on applications of statistics in biology.

Dennis F. Lock is in his third year as a Ph.D. candidate in the Department of Statistics at Iowa State University. He is a Research Assistant working on applications of statistics in agriculture. He has a particular interest in statistics in sports.

---

For more information, contact any of the authors:

| | |
|---|---|
| Robin H. Lock | rlock@stlawu.edu |
| Patti Frazer Lock | plock@stlawu.edu |
| Kari Lock Morgan | kari@stat.duke.edu |
| Eric F. Lock | lock@email.unc.edu |
| Dennis F. Lock | dlock@iastate.edu |

For information about class testing the materials, contact
Shannon Corliss at John Wiley and Sons scorliss@wiley.com