To the Reader:

In 2008, I decided to offer this introductory statistics book on my website. The entire original text is included, along with a couple of class handouts and a brief solutions manual written with Lee Panas. You have my permission to use this material for any not-for-profit educational use. I hope you will find something useful.

Best wishes,

Shelley Rasmussen
Department of Mathematical Sciences
University of Massachusetts/Lowell
Lowell, MA 01854

Shelley_Rasmussen@uml.edu

The book *An Introduction to Statistics with Data Analysis* was originally published by Brooks/Cole Publishing Company, Division of Wadsworth, Inc.

ISBN: 0-534-13578-1

AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

CONTENTS

PART ONE DATA ANALYSIS

CHAPTER 3

CHAPTER 4

CHAPTER 5

PART TWO PROBABILITY

PART THREE STATISTICAL INFERENCE

CHAPTER 14 Inferences About Variances 483

CHAPTER 15 Correlation, Regression, and the Method of Least Squares 512

CHAPTER 16 Inferences About Qualitative (or Categorical) Variables 571

Index of Some Formal Statistical Procedures

Inferences about a measure of central tendency

Procedure	Assumptions	Advantages	Limitations	Section
Large-sample inference about a population mean based on the standard Gaussian distribution	Large random sample	We do not need to know the exact probability distribution of the observations.	Sample size needed for reasonable approximation depends on the shape of the (unknown) distribution of the observations.	10-1
Inferences about a population mean based on a t distribution (t test or one-sample t test)	Random sample of observations from a Gaussian distribution	Procedure tends to be robust to deviations from the Gaussian assumption.	Exact significance levels and confidence levels depend on the Gaussian assumption.	10-3
Inferences about a population mean (or median) based on a Wilcoxon signed rank distribution	Random sample of observations from a continuous, symmetric probability distribution	We do not have to assume that the observations come from a Gaussian probability distribution.	Procedure is not as powerful as the t test if the assumptions for the t test are met.	10-4
Inferences about a population median based on a binomial distribution (sign test)	Random sample of observations from a continuous probability distribution	The only assumption about the population distribution is that it is continuous.	Procedure is much less powerful than those based on a t distribution or a Wilcoxon signed rank distribution, when those procedures are valid.	10-5

Inferences about two measures of central tendency

Procedure	Assumptions	Advantages	Limitations	Section
Large-sample inferences about two means based on the standard Gaussian distribution	Two large, independent random samples	We do not need to know the exact probability distributions of the observations.	Sample sizes needed for reasonable approximations depend on the (unknown) probability distributions of the two samples.	11-1
Inferences about two means based on a t distribution (two-sample t test)	Two independent random samples from Gaussian distributions with the same variance	Inferences tend to be robust to deviations from the Gaussian assumption, and to small deviations from the equal-variance assumption.	Exact significance levels and confidence levels depend on the Gaussian assumption and the equal-variance assumption.	11-3
Inferences about two measures of central tendency based on a Wilcoxon–Mann–Whitney distribution	Two independent random samples; the two distributions have the same shape and variation	We do not have to assume that the observations come from Gaussian distributions.	The procedure is not as powerful as the two-sample t test when the assumptions for the two-sample t test are satisfied.	11-4
Inferences about two medians based on a hypergeometric distribution	Two independent random samples from continuous distributions	We have to assume only that the two population distributions are continuous.	The procedure is much less powerful than those based on a t distribution or a Wilcoxon–Mann–Whitney distribution when those procedures are valid.	11-5

Inferences about two measures of central tendency based on paired samples

Procedure	Assumptions	Advantages	Limitations	Section
Base inferences on the differences between observations within a pair. With these differences, use a procedure for making inferences about a single measure of central tendency (for the distribution of differences).	Random sample of pairs of observations; other assumptions depending on the procedure used (whether based on a *t* distribution, Wilcoxon signed rank distribution, or a binomial distribution)	Advantages of the procedure we select are the same as for the one-sample case.	Limitations of the procedure we select are the same as for the one-sample case.	11-6

Comparing several means in a single-factor experiment

Procedure	Assumptions	Advantages	Limitations	Section
One-way analysis of variance	Independent random samples from Gaussian distributions with the same variance	Procedure tends to be robust to deviations from the Gaussian assumption, and to small deviations from the equal-variance assumption.	Exact probabilities for inferences depend on the Gaussian assumption and the equal-variance assumption.	12-2
Kruskal–Wallis test	Independent random samples from distributions that have the same shape and variation	We do not need to assume Gaussian observations.	This procedure is not as powerful as one-way analysis of variance when that procedure is valid.	12-3

Comparing several means (or treatments) in a randomized block experiment

Procedure	Assumptions	Advantages	Limitations	Section
Parametric analysis, using analysis of variance	Observations are independent, from Gaussian distributions with equal variances; relative treatment effects are the same for each block	Inferences tend to be robust to deviations from the Gaussian assumption, and to small deviations from the equal-variance assumption.	Exact probabilities for inferences depend on all of the assumptions being satisfied.	12-4
Friedman's test	Observations are independent, from distributions with similar shape and variation; relative treatment effects are the same for each block	We do not need to assume Gaussian observations.	Procedure is not as powerful as the parametric analysis when the assumptions for the parametric analysis are satisfied.	12-5

Assessing the effects of two factors upon a response variable

Procedure	Assumptions	Advantages	Limitations	Section
Two-factor analysis of variance	Independent observations from Gaussian distributions with equal variances	Procedure tends to be robust to deviations from the Gaussian assumption, and to small deviations from the equal-variance assumption	Exact probabilities for inferences depend on all of the assumptions being satisfied.	13-1

PREFACE

This book is intended for a one- or two-semester introduction to statistics. The discussion is not calculus-based; the only prerequisite is high school algebra.

The emphasis is on the art of statistical thinking. I believe that a course emphasizing statistical thinking about applied problems ought to be anyone's introduction to statistics, no matter what major or year in college. Everyone should understand the usefulness of statistics in addressing real-world problems. Such an understanding would enrich the lives of all students and motivate some to further study in the theory and application of statistics.

Almost all of the examples and exercises in this book are based on real data sets. In a few cases I felt forced to invent a data set to illustrate an idea, because I did not have a real example at hand. Even then I based the example on a realistic application. As a student and a teacher, I have always appreciated real examples in references. I believe that students will be more motivated to study statistics if its usefulness is immediately apparent. This will be most obvious in a book if examples and exercises illustrate the use of statistics in real investigations.

Data analysis is introduced at the beginning of the book, in Part I, and used throughout. Data analysis involves the use of simple graphical and tabular techniques to gain an understanding of the information in a data set. Regrettably, techniques of data analysis are not familiar to many college graduates, not to mention high school graduates. At a recent multidisciplinary workshop for a select group of exceptional high school teachers (funded by the National Science Foundation and run by the Tsongas Industrial History Center in Lowell, Massachusetts), one social studies teacher did not understand why we would ever want to graph data and several others said they always skipped graphs in textbooks. My response was that they were missing the opportunity to help their students to understand the many graphical presentations of data, some good and some bad, that appear daily in the media.

In data analysis and formal statistical analysis, the more carefully a data set is collected, the more useful information can be derived from it. When we use the ideas of experimental design, we plan a study in order to address the questions of interest as efficiently as possible. A well-designed study often needs very little formal statistical analysis. A poorly designed study may yield little useful information no matter how much we massage the data. The importance of data collection and experimental design is emphasized throughout the book.

In formal statistical analysis, we use a sample of data to make inferences about a larger population. These inferences take the form of probability statements about the population, based on what we see in the sample. (We have to make certain assumptions about the sample in order for these probability statements to make sense; a good experimental design helps to assure the validity of some of these assumptions.) Since probability statements form the basis of formal statistical inference, we have to discuss some probability. I have kept this discussion to a minimum. Part II contains the essential concepts in probability that we need for statistical inference. Two optional sections, Sections 6-4 and 6-6, contain interesting applications of probability that are not used again later. The reader who does not want to cover the median test (Section 11-5) or Fisher's exact test (Section 16-5) can skip the discussion of the hypergeometric probability distributions in Section 7-3.

A number of topics and techniques of formal statistical inference are presented in Part III. Classical analysis that depends on the assumption of Gaussian (normal) data is discussed for each appropriate application. In addition, for many applications I have included one and sometimes two alternatives to the classical analysis. Section 10-4, for instance, discusses nonparametric inferences about a population mean or median, based on ranks; Section 10-5 covers inferences about a population median based on signs; Section 14-4 discusses robust inferences about two or more variances. I think it is important for students to realize that not all data sets follow a Gaussian distribution and that there are straightforward alternatives to the classical analysis for many applications. Readers who want to consider only classical analyses, however, may skip the sections on alternative approaches without loss of continuity.

Many students and friends helped me by providing data sets, reviews, suggestions, and encouragement during the writing of this book. Among them are Paul Catalano, Dennie Clarke-Hundley, Paul Gavelis, Janet LaBonte, Nicole LaVallee, Mary Lundquist, Alex Olsen, Michele Walsh, and Penny Angus Yepez. Miin-Show Chao helped with a number of computer runs. Lee Panas contributed data sets, reviewed chapters, provided useful advice, and solved all the exercises for the solutions manual.

I appreciate the contributions of the many reviewers who patiently read the various versions of the manuscript, each version better than the previous one in large part because of their comments and advice. These reviewers include: Dr. Richard Alo, University of Houston; Professor David Banks, Carnegie-Mellon University; Dr. Lynne Billard, University of Georgia; Dr. Bill Korin, The American University; Professor Robert Lacher, South Dakota State University; Professor Ed Landauer, Clackamas Community College; Ms. Mary Parker, Austin Community College; Professor Robert Schaefer, Miami University; Professor Paul Speckman, University of Missouri; Professor Jeff Spielman, Roanoke Col-

lege; Professor George Terrell, Virginia Polytechnic Institute; and Dr. Cindy van Es, Cornell University. I am grateful to the developmental editor, John Bergez, whose detailed criticisms of my organization and writing style, though punishing, helped me with revisions that made the book much easier to read. I am also grateful to copy editor Susan Reiland for her careful reading of the manuscript and many helpful suggestions.

My thanks to all the investigators cited in the references, from whose work I benefited. My apologies to anyone who should have been cited and was not. My thanks also to all of the teachers, colleagues, and students from whom I have learned probability and statistics, especially Michael Woodroofe of the University of Michigan.

I am grateful most of all to my family for all of their love and support. My mother, Jackie Guernsey; my sister, Pam Grant; my brothers, John Rasmussen and Bill Guernsey; and my mother-in-law, Mary Olsen, provided much encouragement. My children, James, Emily, and Vin were extremely patient and supportive. My husband, Dick Olsen, deserves the most gratitude, for encouraging me to start this project, for providing advice and inspiration, and for giving me the time (the children were all preschoolers when I started) to do it. Thank you very much!

Shelley Rasmussen