

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. Copyright 1992 by Wadsworth, Inc. Copyright 2006 by Shelley Rasmussen. 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

© 1992 by Wadsworth, Inc.

© 2006 by Shelley Rasmussen

Permission is granted by the author for non-for-profit educational use.

[Shelley\\_Rasmussen@uml.edu](mailto:Shelley_Rasmussen@uml.edu)

---

---

## CONTENTS

---

### PART ONE DATA ANALYSIS

---

#### CHAPTER 1

#### **Introduction**

**1**

- 1-1 An Overview of the Book 3
- 1-2 Data Analysis and the World Bank Data Set 5
- 1-3 Questions to Ask Before Starting Data Analysis 10
- 1-4 Using a Computer Statistical Software Package: Minitab 12
- Summary of Chapter 1 13
- Appendix to Chapter 1: The World Bank Indicators 13
- Minitab Appendix for Chapter 1 15

---

#### CHAPTER 2

#### **Studying One Variable at a Time: Lists, Tables, and Plots**

**27**

- 2-1 Lists and Dot Plots 28
- 2-2 Stem-and-Leaf Plots 31
- 2-3 Frequency Tables, Frequency Plots, and Histograms 36
- 2-4 Describing the Shape of a Distribution 43
- 2-5 Quantiles, Box Plots, and Box Graphs 48
- Summary of Chapter 2 53
- Minitab Appendix for Chapter 2 53
- Exercises for Chapter 2 59

---

**CHAPTER 3**

**Studying One Variable at a Time:  
Descriptive Statistics 71**

- 3-1 Measures of Central Tendency 72
- 3-2 Measures of Variation 80
- Summary of Chapter 3 84
- Minitab Appendix for Chapter 3 84
- Exercises for Chapter 3 89

---

**CHAPTER 4**

**Studying Two Variables at a Time 96**

- 4-1 Two-Way Frequency Tables for Studying the Relationship  
Between Two Qualitative Variables 97
- 4-2 Tables and Graphs for Studying the Relationship Between a  
Quantitative Variable and a Qualitative Variable 100
- 4-3 Scatterplots for Studying the Relationship  
Between Two Quantitative Variables 106
- Summary of Chapter 4 114
- Minitab Appendix for Chapter 4 114
- Exercises for Chapter 4 119

---

**CHAPTER 5**

**Studying More Than Two  
Variables at a Time 133**

- 5-1 Multidimensional Frequency Tables  
for Several Qualitative Variables 134
  - 5-2 Scatterplots for Studying Two Quantitative Variables  
Within Levels of a Qualitative Variable 135
  - 5-3 The Scatterplot Matrix for Several  
Quantitative Variables 138
  - 5-4 Displaying a Quantitative Variable by Geographic  
Location: Framed Rectangles on a Map 144
  - 5-5 Effective Graphs 146
  - Summary of Chapter 5 147
  - Minitab Appendix for Chapter 5 147
  - Exercises for Chapter 5 149
-

---

## PART TWO PROBABILITY

---

### CHAPTER 6

#### **Some Ideas in Probability Needed for Statistical Inference** **174**

- 6-1 Probability as Chances 176
- 6-2 Experiment, Outcome, Sample Space, Events 177
- 6-3 Probability Functions 179
- 6-4 The Odds of an Event (Optional) 183
- 6-5 Conditional Probability, Independent  
and Dependent Events 184
- 6-6 Bayes' Rule (Optional) 188
- 6-7 Random Variables 191
- 6-8 Mean, Variance, and Standard Deviation  
of a Finite Random Variable 194
- Summary of Chapter 6 198
- Exercises for Chapter 6 199

---

### CHAPTER 7

#### **Finite Probability Models Based on Counting Techniques** **213**

- 7-1 Permutations and Combinations 214
- 7-2 The Binomial Distributions 219
- 7-3 The Hypergeometric Distributions 224
- Summary of Chapter 7 232
- Minitab Appendix for Chapter 7 232
- Exercises for Chapter 7 234

---

### CHAPTER 8

#### **The Gaussian (Normal) Distributions** **240**

- 8-1 The Gaussian Distributions 243
- 8-2 Approximating a Distribution of Values  
by a Gaussian Distribution 250
- 8-3 The Central Limit Theorem 256
- Summary of Chapter 8 260

Minitab Appendix for Chapter 8 260  
Exercises for Chapter 8 265

**PART THREE**  
**STATISTICAL INFERENCE**

**CHAPTER 9 Basic Ideas in Statistics 274**

9-1 Some Definitions Related to Statistical Inference 276  
9-2 Three Examples 278  
9-3 The General Strategy of Hypothesis Testing 289  
9-4 Some Comments on Hypothesis Testing 292  
9-5 Some Comments on Experimental Design 299  
Summary of Chapter 9 302  
Exercises for Chapter 9 302

**CHAPTER 10 Inferences About a Measure of Central Tendency 307**

10-1 Large-Sample Inference About a Population Mean  
Based on the Standard Gaussian Distribution 309  
10-2 Large-Sample Inference About a Proportion 314  
10-3 Inferences About a Population Mean (or Median)  
Based on a  $t$  Distribution 316  
10-4 Inferences About a Population Mean (or Median)  
Based on a Wilcoxon Signed Rank Distribution 323  
10-5 Inferences About a Population Median  
Based on a Binomial Distribution 331  
Summary of Chapter 10 336  
Minitab Appendix for Chapter 10 337  
Exercises for Chapter 10 342

**CHAPTER 11 Inferences About Two Measures of Central Tendency 348**

11-1 Inferences About Two Means When  
Sample Sizes Are Large 350

- 11-2 Large-Sample Inference About Two Proportions 354
- 11-3 Inferences About Two Measures of Central Tendency  
Based on a  $t$  Distribution 357
- 11-4 Inferences About Two Measures of Central Tendency  
Based on a Wilcoxon–Mann–Whitney Distribution 363
- 11-5 Inferences About Two Medians Based on a  
Hypergeometric Distribution 369
- 11-6 Inferences About Measures of Central Tendency  
Based on Paired Samples 372
- Summary of Chapter 11 377
- Minitab Appendix for Chapter 11 378
- Exercises for Chapter 11 381

**CHAPTER 12****Comparing Several Means: Single-Factor  
and Randomized Block Experiments 399**

- 12-1 Comparing Measures of Central Tendency Two at a Time  
Using the Bonferroni Method 400
- 12-2 Inferences About Several Means in a Single-Factor  
Experiment: One-Way Analysis of Variance 402
- 12-3 Nonparametric Analysis of a Single-Factor Experiment:  
The Kruskal–Wallis Test 411
- 12-4 Parametric Analysis of a Randomized Block Experiment 418
- 12-5 Nonparametric Analysis of a Randomized Block  
Experiment: Friedman’s Test 428
- Summary of Chapter 12 433
- Minitab Appendix for Chapter 12 433
- Exercises for Chapter 12 440

**CHAPTER 13****Two-Factor Experiments: Balanced,  
Completely Randomized, Factorial Designs 451**

- 13-1 Two-Factor Analysis of Variance 454
- 13-2 Two-Factor Experiments with Each Factor at Two Levels 466
- Summary of Chapter 13 473
- Minitab Appendix for Chapter 13 473
- Exercises for Chapter 13 476

---

**CHAPTER 14 Inferences About Variances 483**

- 14-1 Parametric Inferences About a Variance 484
- 14-2 Parametric Inferences About Two Variances 487
- 14-3 Parametric Inferences About More Than Two Variances 490
- 14-4 Robust Inferences About Two or More Variances 494
- Summary of Chapter 14 498
- Minitab Appendix for Chapter 14 498
- Exercises for Chapter 14 502

---

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

- 15-1 The Linear Correlation Coefficient 513
- 15-2 A Parametric Test That a Linear Correlation Coefficient Equals Zero 525
- 15-3 Rank Correlation and a Nonparametric Test for Independence of Two Quantitative Variables 529
- 15-4 Simple Linear Regression and the Method of Least Squares 533
- 15-5 Correlation and Simple Linear Regression 540
- 15-6 A Brief Introduction to Multiple Regression 544
- Summary of Chapter 15 551
- Minitab Appendix for Chapter 15 552
- Exercises for Chapter 15 556

---

**CHAPTER 16 Inferences About Qualitative (or Categorical) Variables 571**

- 16-1 The Chi-Square Goodness-of-Fit Test 573
- 16-2 Small-Sample Inference About a Proportion Based on a Binomial Distribution 576
- 16-3 The Chi-Square Test of Independence of Two Qualitative Variables 579
- 16-4 Comparing the Distribution of a Qualitative Variable Across Populations 583

16-5	Testing for Association in a $2 \times 2$ Frequency Table, Using a Hypergeometric Distribution	586
	Summary of Chapter 16	592
	Minitab Appendix for Chapter 16	593
	Exercises for Chapter 16	595
	Additional Exercises	605
	Numerical Answers to Selected Exercises	625
Appendix 1	The Wilcoxon Signed Rank Distributions	630
Appendix 2	The Wilcoxon–Mann–Whitney Distributions	634
Appendix 3	The Kruskal–Wallis Distributions	637
Appendix 4	Statistical Tables	639
	Glossary of Some Minitab Commands	667
	Bibliography	687
	Subject Index	699
	Source Index to Referenced Examples and Exercises	703

## Index of Some Formal Statistical Procedures

### *Inferences about a measure of central tendency*

Procedure	Assumptions	Advantages	Limitations	Section
Large-sample inferences 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 <i>t</i> distribution ( <i>t</i> test or one-sample <i>t</i> 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 <i>t</i> test if the assumptions for the <i>t</i> 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 <i>t</i> 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 <i>t</i> distribution (two-sample <i>t</i> 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 <i>t</i> test when the assumptions for the two-sample <i>t</i> 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 <i>t</i> 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 <i>t</i> 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

---

---

## P R E F A C E

---

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 Yopez. 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 Woodroffe 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*