The Basic Practice of Statistics

Seventh Edition

MOORE
NOTZ
FLIGNER

The Basic Practice of Statistics

SEVENTH EDITION

The Basic Practice of Statistics

DAVID S. MOORE • WILLIAM I. NOTZ • MICHAEL A. FLIGNER

Purdue University *The Ohio State University* *University of California at Santa Cruz*

W. H. Freeman & Company

A Macmillan Education Imprint

Publisher:	Terri Ward
Senior Acquisitions Editor:	Karen Carson
Marketing Manager:	Cara LeClair
Development Editors:	Leslie Lahr and Jorge Amaral
Associate Editor:	Marie Dripchak
Executive Media Editor:	Laura Judge
Media Editor:	Catriona Kaplan
Associate Media Editor:	Liam Ferguson
Editorial Assistant:	Victoria Garvey
Marketing Assistant:	Bailey James
Photo Editor:	Cecilia Varas
Photo Researcher:	Eileen Liang
Cover and Text Designer:	Vicki Tomaselli
Managing Editor:	Lisa Kinne
Senior Project Manager:	Denise Showers, Aptara®, Inc.
Illustrations and Composition:	Aptara®, Inc.
Production Manager:	Julia DeRosa
Printing and Binding:	RR Donnelley
Cover Credit:	© SoberP/istockphoto

Library of Congress Control Number: 2014950586

Student Edition Hardcover (packaged with EESEE/CrunchIt! access card):
ISBN-13: 978-1-4641-4253-6
ISBN-10: 1-4641-4253-X

Student Edition Loose-leaf (packaged with EESEE/CrunchIt! access card):
ISBN-13: 978-1-4641-7990-7
ISBN-10: 1-4641-7990-5

Instructor Complimentary Copy:
ISBN-13: 978-1-4641-7988-4
ISBN-10: 1-4641-7988-3

Printed in the United States of America

First printing

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010
Houndmills, Basingstoke RG21 6XS, England
www.whfreeman.com

BRIEF CONTENTS

*Starred material is not required for later parts of the text.

CONTENTS

*Starred material is not required for later parts of the text.

TO THE INSTRUCTOR: About This Book

Welcome to the seventh edition of *The Basic Practice of Statistics*. As the name suggests, this text provides an introduction to the practice of statistics that aims to equip students to carry out common statistical procedures and to follow statistical reasoning in their fields of study and in their future employment.

The Basic Practice of Statistics is designed to be accessible to college and university students with limited quantitative background—just "algebra" in the sense of being able to read and use simple equations. It is usable with almost any level of technology for calculating and graphing—from a $15 "two-variable statistics" calculator through a graphing calculator or spreadsheet program through full statistical software. Of course, graphs and calculations are less tedious with good technology, so we recommend making available to your students the most effective technology that circumstances permit.

Despite the lower mathematical level, *The Basic Practice of Statistics* is designed to reflect the actual practice of statistics, where data analysis and design of data production join with probability-based inference to form a coherent science of data. There are good pedagogical reasons for beginning with data analysis (Chapters 1 to 7), then moving to data production (Chapters 8 to 11), and then to probability and inference (Chapters 12 to 27). In studying data analysis, students learn useful skills immediately and get over some of their fear of statistics. Data analysis is a necessary preliminary to inference in practice, because inference requires clean data. Designed data production is the surest foundation for inference, and the deliberate use of chance in random sampling and randomized comparative experiments motivates the study of probability in a course that emphasizes data-oriented statistics. *The Basic Practice of Statistics* gives a full presentation of basic probability and inference (16 of the 27 chapters) but places it in the context of statistics as a whole.

Guiding Principles and the GAISE Guidelines

The Basic Practice of Statistics is based on three principles: balanced content, experience with data, and the importance of ideas. These principles are widely accepted by statisticians concerned about teaching and are directly connected to and reflected by the themes of the College Report of the Guidelines in Assessment and Instruction for Statistics Education (GAISE) Project.

The GAISE Guidelines include six recommendations for the introductory statistics course. The content, coverage, and features of *The Basic Practice of Statistics* are closely aligned to these recommendations:

1. Emphasize statistical literacy and develop statistical thinking. The intent of *The Basic Practice of Statistics* is to be modern *and* accessible. The exposition is straightforward and concentrates on major ideas and skills. One principle of writing for beginners is not to try to tell your students everything you know. Another principle is to offer frequent stopping points, marking off digestible bites of material. Statistical literacy is promoted throughout *The Basic Practice of Statistics* in the many examples and exercises drawn from the popular press and from many fields of study. Statistical thinking is promoted in examples and exercises that give enough background to allow students to consider the meaning of their calculations. Exercises often ask for conclusions that are more than a number (or "reject H_0"). Some exercises require judgment in addition to right-or-wrong calculations and conclusions. Statistics, more than mathematics, depends on judgment for effective use. *The Basic Practice of Statistics* begins to develop students' judgment about statistical studies.

2. Use real data. The study of statistics is supposed to help students work with data in their varied academic disciplines and in their unpredictable later employment. Students learn to work with data by working with data. *The Basic Practice of Statistics* is full of data from many fields of study and from everyday life. Data are more than mere numbers—they are numbers with a context that should play a role in making sense of the numbers and in stating conclusions. Examples and exercises in *The Basic Practice of Statistics,* though intended for beginners, use real data and give enough background to allow students to consider the meaning of their calculations.

3. Stress conceptual understanding rather than mere knowledge of procedures. A first course in statistics introduces many skills, from making a stemplot and calculating a correlation to choosing and carrying out a significance test. In practice (even if not always in the course), calculations and graphs are automated. Moreover, anyone who makes serious use of statistics will need some specific procedures not taught in their college statistics course. *The Basic Practice of Statistics therefore* tries to make clear the larger patterns and big ideas of statistics, not in the abstract, but in the context of learning specific skills and working with specific data. Many of the big ideas are summarized in graphical outlines. Three of the most useful appear inside the front cover. Formulas without guiding principles do students little good once the final exam is past, so it is worth the time to slow down a bit and explain the ideas.

4. Foster active learning in the classroom. Fostering active learning is the business of the teacher, though an emphasis on working with data helps. To this end, we have created interactive applets to our specifications and made them available online. These are designed primarily to help in learning statistics rather than in doing statistics. We suggest using selected applets for classroom demonstrations even if you do not ask students to work with them. *The Correlation and Regression, Confidence Intervals, and P-value of a Test of Significance* applets, for example, convey core ideas more clearly than any amount of chalk and talk.

We also provide web exercises at the end of each chapter. Our intent is to take advantage of the fact that most undergraduates are "web savvy." These exercises require students to search the web for either data or statistical examples and then evaluate what they find. Teachers can use these as classroom activities or assign them as homework projects.

5. Use technology for developing conceptual understanding and analyzing data. Automating calculations increases students' ability to complete problems, reduces their frustration, and helps them concentrate on ideas and problem recognition rather than mechanics. At a minimum, students should have a "two-variable statistics" calculator with functions for correlation and the least-squares regression line as well as for the mean and standard deviation.

Many instructors will take advantage of more elaborate technology, as ASA/MAA and GAISE recommend. And many students who don't use technology in their college statistics course will find themselves using (for example) Excel on the job. *The Basic Practice of Statistics* does not assume or require use of software except in Part V, where the work is otherwise too tedious. It does accommodate software use and tries to convince students that they are gaining knowledge that will enable them to read and use output from almost any source. There are regular "Using Technology" sections throughout the text. Each of these sections displays and comments on output from the same three technologies, representing graphing calculators (the Texas Instruments TI-83 or TI-84), spreadsheets (Microsoft Excel), and statistical software (JMP, Minitab, and CrunchIt!). The output always concerns one of the main teaching examples, so that students can compare text and output.

6. Use assessments to improve and evaluate student learning. Within chapters, a few "Apply Your Knowledge" exercises follow each new idea or skill for a quick check of basic mastery—and also to mark off digestible bites of material. Each of the first four parts of the book ends with a review chapter that includes a point-by-point outline of skills learned, problems students can use to test themselves, and several supplementary exercises. (Instructors can choose to cover any or none of the chapters in Part V, so each of these chapters includes a skills outline.) The review chapters present supplemental exercises without the "I just studied that" context, thus asking for another level of learning. We think it is helpful to assign some supplemental exercises. Many instructors will find that the review chapters appear at the right points for pre-examination review. The "Test Yourself" questions can be used by students to review, self-assess, and prepare for such an examination.

In addition, assessment materials in the form of a test bank and quizzes are available online.

What's New?

The new edition of *The Basic Practice of Statistics* brings many **new examples and exercises**. There are new data sets from a variety of sources, including finance (the relationship between positive articles in the media and the Dow Jones Industrial Average the following week), health (the relationship between salt intake and percent body fat of children), psychology (the relationship between one's attitude about a presidential candidate and how trustworthy the candidate's face appears to be), medicine (the relationship between playing video games and surgical skills), and the environment (global temperatures). Popular examples and exercises such as the Florida manatee regression example return, many with updated data. These are just a few of a large number of new data settings in this edition.

A new edition is also an opportunity to introduce new features and polish the exposition in ways intended to help students learn. Here are some of the changes:

- Each chapter now contains references to online resources to enhance student learning. These include video clips, whiteboard lectures, and technology supplements.
- We have added an introductory chapter, "Getting Started," that instructors may wish to assign to students the first day of classes. This chapter provides an overview of statistical thinking and real examples where the use of statistics can provide valuable insight. It expands on material that was previously included in the Preface, adding motivating examples and exercises.
- Chapter 7 includes descriptions of additional data sets available online that instructors can use for student projects and more extensive data analysis. Along with the description of the data sets, we provide a few suggestions for how they might be used.
- We have added some basic material on resampling and permutation tests in optional sections at the end of Chapters 15, 17, 20, and 21. We hope that instructors who want to introduce students to resampling methods will find this new material useful.
- The essay on data ethics is now Chapter 10, and follows the format of other chapters in the book.
- We have added output from JMP to the "Using Technology" sections.
- The content in Parts I and II has been rewritten to accommodate instructors who prefer to teach data production (Part II) before data exploration (Part I). Instructors can teach these parts in either order while maintaining the continuity of the material.
- Sections are now numbered for easier reference.

FEATURES OF *THE BASIC PRACTICE OF STATISTICS*, Seventh Edition

In this chapter we cover...

Each chapter opener offers a brief overview of where the chapter is heading, often with reference to previous chapters, and includes a section outline of the major topics that will be covered.

In this chapter we cover...

EXAMPLE 2.9 Comparing Graduation Rates

GRADRATE

STATE: Federal law requires all states in the United States to use a common computation of on-time high school graduation rates beginning with the 2010–11 school year. Previously, states chose one of several computation methods that gave answers that could differ by more than 10%. This common computation allows for meaningful comparison of graduation rates between the states.

We know from Table 1.1 (page 22) that the on-time high school graduation rates varied from 59% in the District of Columbia to 88% in Iowa. The U.S. Census Bureau divides the 50 states and the District of Columbia into four geographical regions: the Northeast (NE), Midwest (MW), South (S), and West (W). The region for each state is included in Table 1.1. Do the states in the four regions of the country display distinct distributions of graduation rates? How do the mean graduation rates of the states in each of these regions compare?

PLAN: Use graphs and numerical descriptions to describe and compare the distributions of on-time high school graduation rates of the states in the four regions of the United States.

SOLVE: We might use boxplots to compare the distributions, but stemplots preserve more detail and work well for data sets of these sizes. Figure 2.5 displays the stemplots with the stems lined up for easy comparison. The stems have been split to better display the distributions. The stemplots overlap, and some care is needed when comparing the four stemplots as the sample sizes differ, with some stemplots having more leaves than others. None of the plots shows strong skewness, although the South has one low observation that stands apart from the others with this choice of stems. The states in the Northeast and Midwest have distributions that are similar to each other, as do those in the South and West. The graduation rates tend to be higher for the states in the Northeast and Midwest and more variable for the states in the South and West. With little skewness and no serious outliers, we report $\bar{x}$ and s as our summary measures of center and variability of the distribution of the on-time graduation rates of the states in each region:

Region	Mean	Standard Deviation
Midwest	82.92	4.25
Northeast	82.56	3.47
South	75.93	7.36
West	73.58	6.73

Midwest		Northeast		South		West	
8	66678	8	67	8	66	8	
8	01334	8	33334	8	123	8	002
7	7	7	77	7	5688	7	6668
7	4	7		7	1124	7	4
6		6		6	7	6	88
6		6		6		6	23
5		5		5	9	5	

FIGURE 2.5 Stemplots comparing the distributions of graduation rates for the four census regions from Table 1.1, for Example 2.9.

CONCLUDE: The table of summary statistics confirms what we see in the stemplots. The states in the Midwest and Northeast are quite similar to each other, as are those in the South and West. The states in the Midwest and Northeast have a higher mean graduation rate as well as a smaller standard deviation than those in the South and West. ■

4-Step Examples

In Chapter 2, students learn how to use the four-step process for working through statistical problems: State, Plan, Solve, Conclude. By observing this framework in use in selected examples throughout the text and practicing it in selected exercises, students develop the ability to solve and write reports on real statistical problems encountered outside the classroom.

Apply Your Knowledge

Major concepts are immediately reinforced with problems that are interspersed throughout the chapter (often following examples). These problems allow students to practice their skills concurrently as they work through the text.

Apply Your Knowledge

Photo Researchers/Getty Images

2.10 $\bar{x}$ and s by Hand. Radon is a naturally occurring gas and is the second leading cause of lung cancer in the United States.[9] It comes from the natural breakdown of uranium in the soil and enters buildings through cracks and other holes in the foundations. Found throughout the United States, levels vary considerably from state to state. Several methods can reduce the levels of radon in your home, and the Environmental Protection Agency recommends using one of these if the measured level in your home is above 4 picocuries per liter. Four readings from Franklin County, Ohio, where the county average is 8.4 picocuries per liter, were 6.2, 12.8, 7.6, and 15.4.

(a) Find the mean step-by-step. That is, find the sum of the four observations and divide by 4.

(b) Find the standard deviation step-by-step. That is, find the deviation of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 2.7 shows the method.

(c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain $\bar{x}$ and s. Do the results agree with your hand calculations?

LaunchPad Online Resources

Many sections end with references to the most relevant and helpful online resources (chosen by the authors and available in LaunchPad) for students to use for further explanation or practice.

LaunchPad Online Resources

- The **Snapshots video**, *Summarizing Quantitative Data,* provides an overview of the need for measures of center and variability as well as some details of the computations.
- The **StatClips Examples video**, *Summaries of Quantitative Data Example C,* gives the details for the computation of the mean, median, and standard deviation in a small example. You can verify the computations along with the video, either by hand or using your technology.
- The **StatClips Examples videos**, *Basic Principles of Exploring Data Example B* and *Basic Principles of Exploring Data Example C,* emphasize the need to examine outliers and understand them, rather than simply discarding observations that don't seem to fit.

Using Technology

Located where most appropriate, these special sections display and comment on the output from graphing calculators, spreadsheets, and statistical software in the context of examples from the text.

2.9 Using technology

Although a calculator with "two-variable statistics" functions will do the basic calculations we need, more elaborate tools are helpful. Graphing calculators and computer software will do calculations and make graphs as you command, freeing you to concentrate on choosing the right methods and interpreting your results. Figure 2.4 displays output describing the travel times to work of 20 people in New York State (Example 2.3). Can you find $\bar{x}$, s, and the five-number summary in each output? The big message of this section is: *Once you know what to look for, you can read output from any technological tool.*

The displays in Figure 2.4 come from a Texas Instruments graphing calculator, the Minitab, CrunchIt!, and JMP statistical programs, and the Microsoft Excel spreadsheet program. Minitab and JMP allow you to choose what descriptive measures you want, whereas the descriptive measures in the CrunchIt! output are provided by default. Excel and the calculator give some things we don't need. Just ignore the extras. Excel's "Descriptive Statistics" menu item doesn't give the quartiles. We used the spreadsheet's separate quartile function to get Q_1 and Q_3.

Texas Instruments Graphing Calculator

```
1-Var Stats
 x̄=31.25
 Σx=625
 Σx²=28625
 Sx=21.8773495
 σx=21.32340264
↓n=20
```

```
1-Var Stats
↑n=20
 minX=5
 Q1=15
 Med=22.5
 Q3=42.5
 maxX=85
```

Minitab

Session

Descriptive Statistics: NYtime

```
          Total
variable  Count   Mean  StDev  Variance  Minimum     Q1  Median     Q3  Maximum
NYtime       20  31.25  21.88    478.62     5.00  15.00   22.50  43.75    85.00
```

CrunchIt!

Results - Descriptive Statistics

Export

	n	Sample Mean	Standard Deviation	Min	Q1	Median	Q3	Max
Minutes	20	31.25	21.88	5	15	22.50	42.50	85

Microsoft Excel

Book1

	A	B	C	D
1	*minutes*			
2				
3	Mean	31.25		
4	Standard Error	4.891924064		
5	Median	22.5	QUARTILE(A2:A21,1)	15
6	Mode	15	QUARTILE(A2:A21,3)	42.5
7	Standard Deviation	21.8773495		
8	Sample Variance	478.6184211		
9	Kurtosis	0.329884126		
10	Skewness	1.040110836		
11	Range	80		
12	Minimum	5		
13	Maximum	85		
14	Sum	625		
15	Count	20		

Sheet4 Sheet1 Sheet2 Sheet

JMP Output

Distributions

NYtime

Quantiles

100%	maximum	85
75%	quartile	43.75
50%	median	22.5
25%	quartile	15
0%	minimum	5

Summary Statistics

Mean	31.25
Std Dev	21.877349
N	20

FIGURE 2.4
Output from a graphing calculator, three statistical software packages, and a spreadsheet program describing the data on travel times to work in New York State.

Statistics in Your World

These brief asides in each chapter illustrate major concepts or present cautionary tales through entertaining and relevant stories, allowing students to take a break from the exposition while staying engaged.

HE SAID, SHE SAID.

Height, weight, and body mass distributions in this book come from actual measurements by a government survey. That is a good thing. When *asked* their weight, almost all women say they weigh less than they really do. Heavier men also underreport their weight—but lighter men claim to weigh more than the scale shows. We leave you to ponder the psychology of the two sexes. Just remember that "say-so" is no substitute for measuring.

Chapter Summary and Link It

Each chapter concludes with a summary of the chapter specifics, including major terms and processes, followed by a brief discussion of how the chapter links to material from both previous and upcoming chapters.

CHAPTER 4 SUMMARY

Chapter Specifics

- To study relationships between variables, we must measure the variables on the same group of individuals.
- If we think that a variable x may explain or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable**.
- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph. Always plot the explanatory variable, if there is one, on the x axis of a scatterplot.
- Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.
- In examining a scatterplot, look for an overall pattern showing the **direction**, **form**, and **strength** of the relationship and then for **outliers** or other deviations from this pattern.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).
- **Form: Linear relationships**, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.
- **Strength:** The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
- The **correlation** r measures the direction and strength of the linear association between two quantitative variables x and y. Although you can calculate a correlation for any scatterplot, r measures only straight-line relationships.

Link It

In Chapters 1 to 3, we focused on exploring features of a single variable. In this chapter, we continued our study of exploratory data analysis but for the purpose of examining relationships *between* variables. A useful tool for exploring the relationship between two variables is the scatterplot. When the relationship is linear, correlation is a numerical measure of the strength of the linear relationship.

It is tempting to assume that the patterns we observe in our data hold for values of our variables that we have not observed—in other words, that additional data would continue to conform to these patterns. The process of identifying underlying patterns would seem to assume that this is the case. But is this assumption justified? Parts II to V of the book answer this question.

Check Your Skills and Chapter Exercises

Each chapter ends with a series of multiple-choice problems that test students' understanding of basic concepts and their ability to apply the concepts to real-world statistical situations. The multiple-choice problems are followed by a set of more in-depth exercises that allow students to make judgments and draw conclusions based on real data and real scenarios.

CHECK YOUR SKILLS

4.14 Researchers collect data on 5,134 American adults younger than 60. They measure the reaction times (in seconds) of each subject to a stimulus on a computer screen and how many years later the subject died.[10]

The researchers are interested in whether reaction time can predict time to death (in years). When you make a scatterplot, the explanatory variable on the x axis

(a) is the reaction time.
(b) is the time to death.
(c) can be either reaction time or time to death.

4.15 The researchers in Exercise 4.14 found that people with slower reaction times tended to die sooner. In a scatterplot of the reaction time and the number of years to death, you expect to see

(a) a positive association.
(b) very little association.
(c) a negative association.

4.16 Figure 4.7 is a scatterplot of school GPA against IQ test scores for 15 seventh-grade students. There is one low outlier in the plot. The IQ and GPA scores for this student are

(a) IQ = 0.5, GPA = 103.
(b) IQ = 103, GPA = 0.5.
(c) IQ = 103, GPA = 7.6.

4.17 If we leave out the low outlier, the correlation for the remaining 14 points in Figure 4.7 is closest to

(a) 0.9. (b) −0.9. (c) 0.1.

4.18 What are all the values that a correlation r can possibly take?

(a) $r \geq 0$ (b) $0 \leq r \leq 1$ (c) $-1 \leq r \leq 1$

4.19 If the correlation between two variables is close to 0, you can conclude that a scatterplot would show

(a) a strong straight-line pattern.
(b) a cloud of points with no visible pattern.
(c) no straight-line pattern, but there might be a strong pattern of another form.

4.20 The points on a scatterplot lie very close to a straight line. The correlation between x and y is close to

(a) −1. (b) 1. (c) either −1 or 1, we can't say which.

4.21 A statistics professor warns her class that her second midterm is always harder than the first. She tells her class that students always score 10 points worse on the second midterm compared to their score on the first midterm. This means that the correlation between students' scores on the first and second exam is

(a) 1. (b) −1. (c) Can't tell without seeing the data.

4.22 Researchers asked mothers how much soda (in ounces) their kids drank in a typical day. They also asked these mothers to rate how aggressive their kids were on a scale of 1 to 10, with larger values corresponding to a greater degree of aggression.[11] The correlation between amount of soda consumed and aggression rating was found to be $r = 0.3$. If the researchers had measured amount of soda consumed in liters instead

CHAPTER 4 EXERCISES

4.24 Scores at the Masters. The Masters is one of the four major golf tournaments. Figure 4.8 is a scatterplot of the scores for the first two rounds of the 2013 Masters for all the golfers entered. Only the 60 golfers with the lowest two-round total advance to the final two rounds (unless several people are tied for 60th place, in which case all those tied for 60th place advance). The plot has a grid pattern because golf scores must be whole numbers.[13] MASTR13

(a) Read the graph: What was the lowest score in the first round of play? How many golfers had this low score? What were their scores in the second round?
(b) Read the graph: Alan Dunbar had the highest score in the second round. What was this score? What was Dunbar's score in the first round?

4.25 Happy states. Human happiness or well-being can be assessed either subjectively or objectively. Subjective assessment can be accomplished by listening to what people say. Objective assessment can be made from data related to well-being such as income, climate, availability of entertainment, housing prices, lack of traffic congestion, and so on. Do subjective and objective assessments agree? To study this, investigators made both subjective and objective assessments of happiness for each of the 50 states. The subjective measurement was the mean score on a life-satisfaction question found on the Behavioral Risk Factor Surveillance System (BRFSS), which is a state-based system of health surveys. Lower scores indicate a greater degree of happiness. To objectively assess happiness, the investigators computed a mean well-being score (called

Web Exercises

A final set of exercises asks students to investigate data and statistical issues by researching topics online. These exercises tend to be more involved and provide an opportunity for students to dig deep into contemporary issues and special applications of statistics.

Exploring the Web

6.34 Promoting women. In academics, faculty typically start as assistant professors, are promoted to associate professor (and gain tenure), and finally reach the rank of full professor. Some have argued that women have a harder time gaining promotion to associate and full professor than do men. Do data support this argument? Search the web to find the number of faculty by rank and gender at some university. Do you see a pattern that suggests that the proportion of women decreases as rank increases? We found several sources of data by doing a Google search on "faculty head count by rank and gender." In addition to discussing the pattern you find, provide the data, the name of the school, and the source of the data.

6.35 Accidental deaths and age. Accidental deaths are shocking and tragic. Do the ways in which people die by accident change with age? Look at the most recent *Statistical Abstract of the United States* (`www.census.gov/compendia/statab/`) and make a two-way table that provides the counts of deaths due to accidents from various causes for three different age groups. What do you conclude?

6.36 Simpson's paradox. Find an example of Simpson's paradox and discuss how your example illustrates the paradox. Two examples that we found (thanks to Patricia Humphrey at Georgia Southern University) are `www.nytimes.com/2006/07/15/education/15report.html` and `online.wsj.com/article/SB125970744553071829.html`.

Online Data for Additional Analyses

References to larger data sets are suggested in Chapter 7 to provide an opportunity for students to apply the methods of Chapters 1–6 to explore data on their own. This is intended to reinforce the idea of exploratory data analysis as a tool for exploring data.

Online Data for Additional Analyses

1. SAT, ACT, and teacher salaries for 2013 for each of the 50 states and the District of Columbia are available in the data set SATACT. One could use these data to carry out analyses for ACT scores similar to those for the SAT scores in Chapters 5 and 6. For example, repeat the analyses in Exercises 5.50 (page 157) and 5.51 (page 158) using ACT scores instead of SAT scores. SATACT

2. The data set MLB contains hitting, pitching, fielding, salary, and win–loss performance data from the 2013 season for all major league baseball teams. These data can be used to determine the correlation between payroll and winning percentage. One can also explore what variables are most highly correlated with winning percentage, and whether variables that measure pitching performance are more highly correlated with winning percentage than variables that measure hitting performance. For example, calculate the correlation between winning percentage and number of home runs, between winning percentage and batting average, between winning percentage and ERA, between winning percentage and strikeouts by pitchers, and between winning percentage and payroll. Which has the highest correlation? These data are from `http://www.baseball-reference.com/`. Visit this website for definitions of several of the variables in the data set. MLB

3. Historical temperature data and whether Punxsutawney Phil saw his shadow are available in the data set PHIL. Repeat the analysis in Exercise 6.32 (page 177), but define what constitutes "six more weeks of winter-like weather" differently. For example, you might decide there were six more weeks of winter-like weather if average temperatures for March were at least one degree below historical averages. PHIL

4. Data from the Ohio Department of Health website are available in the pdf "2013OHH Detail Tables." This is a source of many tables that can be used for further analyses using methods discussed in Chapter 6. For example, conduct an analysis like that in Exercise 7.52 to investigate the relationship between sex and strategies about weight (Question 67 in the Tables). HEALTH

5. The data set WHAT contains three variables and 3848 observations on each. At one time, this was considered a large data set and difficult to explore with software. Use various exploratory methods available in software packages such as JMP and Minitab to find the "hidden pattern" in these data. WHAT

Why Did You Do That?

There is no single best way to organize our presentation of statistics to beginners. That said, our choices reflect thinking about both content and pedagogy. Here are comments on several "frequently asked questions" about the order and selection of material in *The Basic Practice of Statistics.*

- **Why does the distinction between population and sample not appear in Part I?** There is more to statistics than inference. In fact, statistical inference is appropriate only in rather special circumstances. The chapters in Part I present tools and tactics for describing data—any data. These tools and tactics do not depend on the idea of inference from sample to population. Many data sets in these chapters (for example, the several sets of data about the 50 states) do not lend themselves to inference because they represent an entire population. John Tukey of Bell Labs and Princeton, the philosopher of modern data analysis, insisted that the population–sample distinction be avoided when it is not relevant. He used the word "batch" for data sets in general. We see no need for a special word, but we think Tukey was right.

- **Why not begin with data production?** We prefer to begin with data exploration (Part I), as most students will use statistics mainly in settings other than planned research studies in their future employment. We place the design of data production (Part II) after data analysis to emphasize that data-analytic techniques apply to any data. However, it is equally reasonable to begin with data production—the natural flow of a planned study is from design to data analysis to inference. Because instructors have strong and differing opinions on this question, these two topics are now the first two parts of the book, with the text written so that it may be started with either Part I or Part II while maintaining the continuity of the material.

- **Why do Normal distributions appear in Part I?** Density curves such as the Normal curves are just another tool to describe the distribution of a quantitative variable, along with stemplots, histograms, and boxplots. Professional statistical software offers to make density curves from data just as it offers histograms. We prefer not to suggest that this material is essentially tied to probability, as the traditional order does. And we find it helpful to break up the indigestible lump of probability that troubles students so much. Meeting Normal distributions early does this and strengthens the "probability distributions are like data distributions" way of approaching probability.

- **Why not delay correlation and regression until late in the course, as was traditional?** *The Basic Practice of Statistics* begins by offering experience working with data and gives a conceptual structure for this nonmathematical but essential part of statistics. Students profit from more experience with data and from seeing the conceptual structure worked out in relations among variables as well as in describing single-variable data. Correlation and least-squares regression are very important descriptive tools and are often used in settings where there is no population–sample distinction, such as studies of all of a firm's employees. Perhaps most important, the approach taken by *The Basic Practice of Statistics* asks students to think about what kind of relationship lies behind the data (confounding, lurking variables, association doesn't imply causation, and so on), without overwhelming them with the demands of formal inference methods. Inference in the correlation and regression setting is a bit complex, demands software, and often comes right at the end of the course. We find that delaying all mention of correlation and regression to that point means that students often don't master the basic uses and properties of these methods. We consider Chapters 4 and 5 (correlation and regression) essential and Chapter 26 (regression inference) optional.

- **Why use the z procedures for a population mean to introduce the reasoning of inference?** This is a pedagogical issue, not a question of statistics in practice. The two most popular choices for introducing inference are z for a mean and z for a proportion. (Another option is resampling and permutation tests. We have included material on these topics, but have not used them to introduce inference.)

We find z for means quite accessible to students. Positively, we can say up front that we are going to explore the reasoning of inference in the overly simple setting described in the box on page 374 titled Simple Conditions for Inference about a Mean. As this box suggests, exactly Normal population and true simple random sample are as unrealistic as known σ. All the issues of practice—robustness against lack of Normality and application when the data aren't an SRS as well as the need to estimate σ—are put off until, with the reasoning in hand, we discuss the practically useful t procedures. This separation of initial reasoning from messier practice works well.

Negatively, starting with inference for p introduces many side issues: no exact Normal sampling distribution, but a Normal approximation to a discrete distribution; use of $\hat{p}$ in both the numerator and denominator of the test statistic to estimate both the parameter p and $\hat{p}$'s own standard deviation; loss of the direct link between test and confidence interval; and the need to avoid small and moderate sample sizes because the Normal approximation for the test is quite unreliable.

There are advantages to starting with inference for p. Starting with z for means takes a fair amount of time and the ideas need to be rehashed with the introduction of the t procedures. Many instructors face pressure from client departments to cover a large amount of material in a single semester. Eliminating coverage of the "unrealistic" z for means with known variance enables instructors to cover additional, more realistic applications of inference. Also, many instructors believe that proportions are simpler and more familiar to students than means. For instructors who would prefer to introduce inference with z for a proportion, we recommend our book, *Statistics in Practice*.

- **Why didn't you cover Topic X?** Introductory texts ought not to be encyclopedic. We chose topics on two grounds: they are the most commonly used in practice, and they are suitable vehicles for learning broader statistical ideas. Students who have completed the core of the book, Chapters 1 to 12 and 15 to 24, will have little difficulty moving on to more elaborate methods. Chapters 25 to 27 offer a choice of slightly more advanced topics, as do the four companion chapters available online.

ACKNOWLEDGMENTS

We have enjoyed the opportunity to once again rethink how to help beginning students achieve a practical grasp of basic statistics. What students actually learn is not identical to what we teachers think we have "covered," so the virtues of concentrating on the essentials are considerable. We hope that the new edition of *The Basic Practice of Statistics* offers a mix of concrete skills and clearly explained concepts that will help many teachers guide their students toward useful knowledge.

We are grateful to colleagues from two-year and four-year colleges and universities who commented on *The Basic Practice of Statistics:*

Faran Ali, *Simon Fraser University*
Michael Allen, *Glendale Community College*
Paul Lawrence Baker, *Catawba College*
Brigitte Baldi, *University of California—Irvine*
Barbara A. Barnet, *University of Wisconsin—Platteville*
Paul R. Bedard, *Saint Clair Community College*
Marjorie E. Bond, *Monmouth College*
Ryan Botts, *Point Loma Nazarene University*
Mine Cetinkaya-Rundel, *Duke University*
Gary Cochell, *Culver-Stockton College*
Patti Collings, *Brigham Young University*
Phyllis Curtis, *Grand Valley State University*
Carolyn Pillers Dobler, *Gustavus Adolphus College*
John Daniel Draper, *The Ohio State University*
Michelle Everson, *The Ohio State University*
Diane G. Fisher, *University of Louisiana at Lafayette*
James Gray, *University of Washington*
Ellen Gundlach, *Purdue University*
James A. Harding, *Green Mountain College*
James Hartman, *The College of Wooster*
Pat Humphrey, *Georgia Southern University*
Dick Jardine, *Keene State College*
Robert W. Jernigan, *American University*
Jennifer Kaplan, *University of Georgia*
Daniel L. King, *Sarah Lawrence College*
William "Sonny" Kirby, *Gadsden State Community College*
Brian Knaeble, *University of Wisconsin—Stout*
Allyn Leon, *Imperial Valley College*
Karen P. Lundberg, *Colorado State University—Pueblo*
Dana E. Madison, *Clarion University of Pennsylvania*
Kimberly Massaro, *University of Texas at San Antonio*
Jackie Miller, *University of Michigan*
Juliann Moore, *Oregon State University*
Penny Ann Morris, *Polk State College*
Kathleen Mowers, *Owensboro Community and Technical College*
Julia Ann Norton, *California State University—East Bay*
Mary R. Parker, *Austin Community College*
Michael Price, *University of Oregon*
David Rangel, *Bellingham Technical College*
Shane Patrick Redmond, *Eastern Kentucky University*
Scott J. Richter, *University of North Carolina at Greensboro*
Laurence David Robinson, *Ohio Northern University*
Caroline Schruth, *Tacoma Community College*
Mack Shelley, *Iowa State University*
Therese N. Shelton, *Southwestern University*
Haskell Sie, *Pennsylvania State University*
Murray H. Siegel, *Arizona State University—Polytechnic Campus*
Sean Simpson, *Westchester Community College*
Robb Sinn, *University of North Georgia*
Karen H. Smith, *University of West Georgia*
Stephen R. Soltys, *Elizabethtown College*
James Stamey, *Baylor University*
Jeanette M. Szwec, *Cape Fear Community College*
Ramin Vakilian, *California State University—Northridge*
Asokan Mulayath Variyath, *Memorial University of Newfoundland*
Lianwen Wang, *University of Central Missouri*
Barbara B. Ward, *Belmont University*
Yajni Warnapala, *Roger Williams University*
Robert E. White, *Allan Hancock College*
Ronald L. White, *Norfolk State University*
Rachelle Curtis Wilkinson, *Austin Community College*

We extend our appreciation to Ruth Baruth, Terri Ward, Karen Carson, Leslie Lahr, Jorge Amaral, Marie Dripchak, Laura Judge, Catriona Kaplan, Liam Ferguson, Victoria Garvey, Cara LeClair, Bailey James, Cecilia Varas, Eileen Liang, Lisa Kinne, Julia DeRosa, Laurel Sparrow, and other publishing professionals who have contributed to the development, production, and cohesiveness of this book and its online resources.

Special thanks are due to Vicki Tomaselli, whose talents were poured into the aesthetic appeal of this book. We extend our appreciation to Denise Showers of Aptara, Inc., who has offered her knowledge, expertise, and patience tirelessly throughout the production process.

We are deeply indebted to our colleagues, Jackie B. Miller and Patricia B. Humphrey, for their many contributions, insights, time, and humor. Their wisdom and experience in the classroom have added to a level of quality that students and instructors alike have come to expect. Each of them brought to the project their individual strengths and talents, but they did so in the spirit of true teamwork and collaboration.

We would also like to specially thank the authors and reviewers of the supplementary materials available with *The Basic Practice of Statistics,* 7e; their work and dedication to quality have resulted in a robust package of resources that complement the ideas and concepts presented in the text:

Solutions manuals written by Pat Humphrey, *Georgia Southern University*
Solutions accuracy reviewed by Jackie Miller, *University of Michigan*
Test bank written by Christiana Drake, *University of California–Davis*
Test bank accuracy reviewed by Catherine Matos, *Clayton State University*
iClicker slides created by Dilshod Achilov, *Tennessee State University*
iClicker slides accuracy reviewed by Jun Ye, *The University of Akron*
Practice Quizzes written by Leslie Hendrix, *University of South Carolina*
Practice Quizzes accuracy reviewed by Jun Ye, *The University of Akron*
Lecture PowerPoints created by Mark Gebert, *University of Kentucky, Lexington*

The team of statistics educators who created the new StatBoards videos deserve our praise and thanks; their creative works offer intuitive approaches to the key concepts in the course:

Doug Tyson, *Central York High School*
Michelle Everson, *The Ohio State University*
Marian Frazier, *Gustavus Adolphus College*
Aimee Schwab, *University of Nebraska–Lincoln*

Finally, we are indebted to the many statistics teachers with whom we have discussed the teaching of our subject over many years; to people from diverse fields with whom we have worked to understand data; and especially to students whose compliments and complaints have changed and improved our teaching. Working with teachers, colleagues in other disciplines, and students constantly reminds us of the importance of hands-on experience with data and of statistical thinking in an era when computer routines quickly handle statistical details.

David S. Moore, William I. Notz, and Michael A. Fligner

MEDIA AND SUPPLEMENTS

LaunchPad

W. H. Freeman's new online homework system, **LaunchPad,** offers our quality content curated and organized for easy assignability in a simple but powerful interface. We've taken what we've learned from thousands of instructors and hundreds of thousands of students to create a new generation of W. H. Freeman/Macmillan technology.

Curated Units. Combining a curated collection of videos, homework sets, tutorials, applets, and e-Book content, LaunchPad's interactive units give instructors a building block to use as is or as a starting point for customized learning units. A majority of exercises from the text can be assigned as online homework, including an abundance of algorithmic exercises. An entire unit's worth of work can be assigned in seconds, drastically reducing the amount of time it takes for instructors to have their course up and running.

Easily customizable. Instructors can customize the LaunchPad units by adding quizzes and other activities from our vast wealth of resources. They can also add a discussion board, a dropbox, and RSS feed, with a few clicks. LaunchPad allows instructors to customize students' experience as much or as little as desired.

Useful analytics. The gradebook quickly and easily allows instructors to look up performance metrics for classes, individual students, and individual assignments.

Intuitive interface and design. The student experience is simplified. Students' navigation options and expectations are clearly laid out at all times, ensuring they can never get lost in the system.

Assets integrated into LaunchPad include the following:

Interactive e-Book. Every LaunchPad e-Book comes with powerful study tools for students, video and multimedia content, and easy customization for instructors. Students can search, highlight, and bookmark, making it easier to study and access key content. And teachers can ensure that their classes get just the book they want to deliver: customize and rearrange chapters, add and share notes and discussions, and link to quizzes, activities, and other resources.

LEARNINGCurve **LearningCurve** provides students and instructors with powerful adaptive quizzing, a game-like format, direct links to the e-Book, and instant feedback. The quizzing system features questions tailored specifically to the text and adapts to students' responses, providing material at different difficulty levels and topics based on student performance.

FREEMAN SolutionMaster **SolutionMaster** offers an easy-to-use web-based version of the instructor's solutions, allowing instructors to generate a solution file for any set of homework exercises.

New **StatBoards videos** are brief whiteboard videos that illustrate difficult topics through additional examples, written and explained by a select group of statistics educators.

New Stepped Tutorials are centered on algorithmically generated quizzing with step-by-step feedback to help students work their way toward the correct solution. These new exercise tutorials (two to three per chapter) are easily assignable and assessable.

Statistical Video Series consists of StatClips, StatClips Examples, and Statistically Speaking "Snapshots." View animated lecture videos, whiteboard lessons, and documentary-style footage that illustrate key statistical concepts and help students visualize statistics in real-world scenarios.

New Video Technology Manuals available for TI-83/84 calculators, Minitab, Excel, JMP, SPSS, R, Rcmdr, and CrunchIt!® provide brief instructions for using specific statistical software.

Updated StatTutor Tutorials offer multimedia tutorials that explore important concepts and procedures in a presentation that combines video, audio, and interactive features. The newly revised format includes built-in, assignable assessments and a bright new interface.

Updated Statistical Applets give students hands-on opportunities to familiarize themselves with important statistical concepts and procedures, in an interactive setting that allows them to manipulate variables and see the results graphically. Icons in the textbook indicate when an applet is available for the material being covered.

CrunchIt!® is W. H. Freeman's web-based statistical software that allows users to perform all the statistical operations and graphing needed for an introductory statistics course and more. It saves users time by automatically loading data from BPS, and it provides the flexibility to edit and import additional data.

JMP Student Edition (developed by SAS) is easy to learn and contains all the capabilities required for introductory statistics, including pre-loaded data sets from BPS. JMP is the leading commercial data analysis software of choice for scientists, engineers, and analysts at companies throughout the globe (for Windows and Mac).

Stats@Work Simulations put students in the role of the statistical consultant, helping them better understand statistics interactively within the context of real-life scenarios.

***EESEE* Case Studies** (*Electronic Encyclopedia of Statistical Examples and Exercises*), developed by The Ohio State University Statistics Department, teach students to apply their statistical skills by exploring actual case studies using real data.

Data files are available in CrunchIt!, JMP, ASCII, Excel, TI, Minitab, and SPSS (an IBM Company)* formats.

Student Solutions Manual provides solutions to the odd-numbered exercises in the text. It is available electronically within LaunchPad, as well as in print form.

Interactive Table Reader allows students to use statistical tables interactively to seek the information they need.

Instructor's Guide with Full Solutions includes teaching suggestions, chapter comments, and detailed solutions to all exercises. It is available electronically within LaunchPad.

*SPSS was acquired by IBM in October 2009.

Test Bank offers hundreds of multiple-choice questions. It is also available on CD-ROM (for Windows and Mac), where questions can be downloaded, edited, and resequenced to suit each instructor's needs.

Lecture PowerPoint Slides offer a customizable, detailed lecture presentation of statistical concepts covered in each chapter of BPS.

Additional Resources Available with *BPS*

Companion Website www.whfreeman.com/bps7e This open-access website includes statistical applets, data files, supplementary exercises, and self-quizzes. The website also offers companion chapters covering nonparametric tests, multiple regression, further topics in ANOVA, and statistics for quality control and capability. Instructor access to the Companion Website requires user registration as an instructor and features all the open-access student web materials, plus:

- **Instructor's Guide with Full Solutions**
- **Test Bank**
- **Lecture PowerPoint Slides containing all textbook figures and tables**
- **Instructor version of EESEE with solutions to the exercises in the student version**

Special Software Packages Student versions of JMP and Minitab are available for packaging with the text. JMP is available inside LaunchPad at no additional cost. Contact your W. H. Freeman representative for information or visit www.whfreeman.com.

Course Management Systems W. H. Freeman and Company provides courses for Blackboard, Angel, Desire2Learn, Canvas, Moodle, and Sakai course management systems. These are completely integrated solutions that instructors can customize and adapt to meet teaching goals and course objectives. Visit macmillanhighered.com/Catalog/other/Coursepack for more information.

i-clicker **i-clicker** is a two-way radio-frequency classroom response solution developed by educators for educators. Each step of i-clicker's development has been informed by teaching and learning. To learn more about packaging i-clicker with this textbook, please contact your local sales rep or visit www1.iclicker.com.

ABOUT THE AUTHORS

David S. Moore is Shanti S. Gupta Distinguished Professor of Statistics, Emeritus, at Purdue University and was the 1998 president of the American Statistical Association. He received his A.B. from Princeton and his Ph.D. from Cornell, both in mathematics. He has written many research papers in statistical theory and served on the editorial boards of several major journals. Professor Moore is an elected fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He has served as program director for statistics and probability at the National Science Foundation. Professor Moore has made many contributions to the teaching of statistics. He was the content developer for the Annenberg/Corporation for Public Broadcasting college-level telecourse Against All Odds: Inside Statistics and for the series of video modules Statistics: Decisions through Data, intended to aid the teaching of statistics in schools. He is the author of influential articles on statistics education and of several leading texts. Professor Moore has served as president of the International Association for Statistical Education and has received the Mathematical Association of Americas national award for distinguished college or university teaching of mathematics.

William I. Notz is Professor of Statistics at the Ohio State University. He received his B.S. in physics from the Johns Hopkins University and his Ph.D. in mathematics from Cornell University. His first academic job was as an assistant professor in the Department of Statistics at Purdue University. While there, he taught the introductory concepts course with Professor Moore and as a result of this experience he developed an interest in statistical education. Professor Notz is a co-author of *EESEE* (the *Electronic Encyclopedia of Statistical Examples and Exercises*) and co-author of *Statistics: Concepts and Controversies.*

Professor Notz's research interests have focused on experimental design and computer experiments. He is the author of several research papers and of a book on the design and analysis of computer experiments. He is an elected fellow of the American Statistical Association. He has served as the editor of the journal *Technometrics* and as editor of the *Journal of Statistics Education*. He has served as the Director of the Statistical Consulting Service, as acting chair of the Department of Statistics for a year, and as an Associate Dean in the College of Mathematical and Physical Sciences at the Ohio State University. He is a winner of the Ohio State University's Alumni Distinguished Teaching Award.

Michael A. Fligner is an Adjunct Professor at the University of California at Santa Cruz and a nonresident Professor Emeritus at the Ohio State University. He received his B.S. in mathematics from the State University of New York at Stony Brook and his Ph.D. from the University of Connecticut. He spent most of his professional career at the Ohio State University where he was vice-chair of the department for over 10 years and also served as Director of the Statistical Consulting Service. He has done consulting work with several large corporations in Central Ohio.

Professor Fligner's research interests are in nonparametric statistical methods and he received the Statistics in Chemistry award from the American Statistical Association for work on detecting biologically active compounds. He is co-author of the book *Statistical Methods for Behavioral Ecology* and received a Fulbright scholarship under the American Republics Research program to work at the Charles Darwin Research Station in the Galapagos Islands. He has been an Associate Editor of the *Journal of Statistical Education*. Professor Fligner is currently associated with the Center for Statistical Analysis in the Social Sciences at the University of California at Santa Cruz.

Ryan Etter/Getty Images

CHAPTER 0

Getting Started

In this chapter we cover...

What's hot in popular music this week? SoundScan knows. SoundScan collects data electronically from the cash registers in more than 14,000 retail outlets and also collects data on download sales from websites. When you buy a CD or download a digital track, the checkout scanner or website is probably telling SoundScan what you bought. SoundScan provides this information to *Billboard* magazine, MTV, and VH1, as well as to record companies and artists' agents.

Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said "Yes." In particular, women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits. But in 2002, the National Institutes of Health declared these findings wrong. Use of hormones after menopause immediately plummeted. Both recommendations were based on extensive studies. What happened?

Is the climate warming? Is it becoming more extreme? An overwhelming majority of scientists now agree that the earth is undergoing major changes in climate. Enormous quantities of data are continuously being collected from weather stations, satellites, and other sources to monitor factors such as the surface temperature on land and sea, precipitation, solar activity, and the chemical composition of air and water. Climate models incorporate this information to make projections of future climate change and can help us understand the effectiveness of proposed solutions.

SoundScan, medical studies, and climate research all produce data (numerical facts), and lots of them. Using data effectively is a large and growing part of most professions, and reacting to data is part of everyday life. In fact, we define statistics as **the science of learning from data.**

Although data are numbers, they are not "just numbers." *Data are numbers with a context.* The number 8.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 8.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our background knowledge and allows us to make judgments. We know that a baby weighing 8.5 pounds is a little above average, and that a human baby is unlikely to weigh 8.5 ounces or 8.5 kilograms (over 18 pounds). The context makes the number informative.

To gain insight from data, we make graphs and do calculations. But graphs and calculations are guided by ways of thinking that amount to educated common sense. Let's begin our study of statistics with an informal look at some aspects of statistical thinking.[1]

0.1 Where the data comes from matters

Although, data can be collected in a variety of ways, the type of conclusion that can be reached from the data depends on how the data were obtained. *Observational studies* and *experiments* are two common methods for collecting data. Let's take a closer look at the hormone replacement data to understand the differences.

EXAMPLE 0.1 Hormone Replacement Therapy

What's behind the flip-flop in the advice offered to women about hormone replacement? The evidence in favor of hormone replacement came from a number of observational studies that compared women who were taking hormones with others who were not. But women who choose to take hormones are very different from women who do not: they are richer and better educated and see doctors more often. These women do many things to maintain their health. It isn't surprising that they have fewer heart attacks.

Large and careful observational studies are expensive, but they are easier to arrange than careful experiments. Experiments don't let women decide what to do. They assign women either to hormone replacement or to dummy pills that look and taste the same as the hormone pills. The assignment is done by a coin toss, so that all kinds of women are equally likely to get either treatment. Part of the difficulty of a good experiment is persuading women to accept the result—invisible to them—of the coin toss. By 2002, several experiments agreed that hormone replacement does *not* reduce the risk of heart attacks, at least for older women. Faced with this better evidence, medical authorities changed their recommendations.[2] ■

Women who chose hormone replacement after menopause were on the average richer and better educated than those who didn't. No wonder they had fewer heart attacks. We can't conclude that hormone replacement reduces heart attacks just because we see this relationship in data. In this example, education and affluence are background factors that help explain the relationship between hormone replacement and good health.

Children who play soccer do better in school (on the average) than children who don't play soccer. Does this mean that playing soccer increases school grades?

Children who play soccer tend to have prosperous and well-educated parents. Once again, education and affluence are background factors that help explain the relationship between soccer and good grades.

Almost all relationships between two observed characteristics or "variables" are influenced by other variables lurking in the background. To understand the relationship between two variables, you must often look at other variables. Careful statistical studies try to think of and measure possible *lurking variables* in order to correct for their influence. As the hormone saga illustrates, this doesn't always work well. News reports often just ignore possible lurking variables that might ruin a good headline like "Playing soccer can improve your grades." The habit of asking, "What might lie behind this relationship?" is part of thinking statistically.

Of course, observational studies are still quite useful. We can learn from observational studies how chimpanzees behave in the wild or which popular songs sold best last week or what percent of workers were unemployed last month. SoundScan's data on popular music and the government's data on employment and unemployment come from *sample surveys,* an important kind of observational study that chooses a part (the sample) to represent a larger whole. Opinion polls interview perhaps 1000 of the 235 million adults in the United States to report the public's views on current issues. Can we trust the results? We'll see that this isn't a simple yes-or-no question. Let's just say that the government's unemployment rate is much more trustworthy than opinion poll results, and not just because the Bureau of Labor Statistics interviews 60,000 people rather than 1000. We can, however, say right away that some samples *can't* be trusted. Consider the following write-in poll.

EXAMPLE 0.2 Would You Have Children Again?

The advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" A few weeks later, her column was headlined "70% OF PARENTS SAY KIDS NOT WORTH IT." Indeed, 70% of the nearly 10,000 parents who wrote in said they would not have children if they could make the choice again. Those 10,000 parents were upset enough with their children to write Ann Landers. Most parents are happy with their kids and don't bother to write. ■

Statistically designed samples, even opinion polls, don't let people choose themselves for the sample. They interview people selected by impersonal chance so that everyone has an equal opportunity to be in the sample. Such a poll showed that 91% of parents *would* have children again. *Where data come from matters a lot.* If you are careless about how you get your data, you may announce 70% "no" when the truth is close to 90% "yes." Understanding the importance of where data come and its relationship to the conclusions that can be reached is an important part of learning to think statistically.

0.2 Always look at the data

Yogi Berra, the Hall of Fame New York Yankee, said it: "You can observe a lot by just watching." That's a motto for learning from data. *A few carefully chosen graphs are often more instructive than great piles of numbers.* Consider the outcome of the 2000 presidential election in Florida.

EXAMPLE 0.3 Palm Beach County

Elections don't come much closer: after much recounting, state officials declared that George Bush had carried Florida by 537 votes out of almost 6 million votes cast. Florida's vote decided the 2000 presidential election and made George Bush, rather than Al Gore, president. Let's look at some data. Figure 0.1 displays a graph that plots votes for the third-party candidate Pat Buchanan against votes for the Democratic candidate Al Gore in Florida's 67 counties.

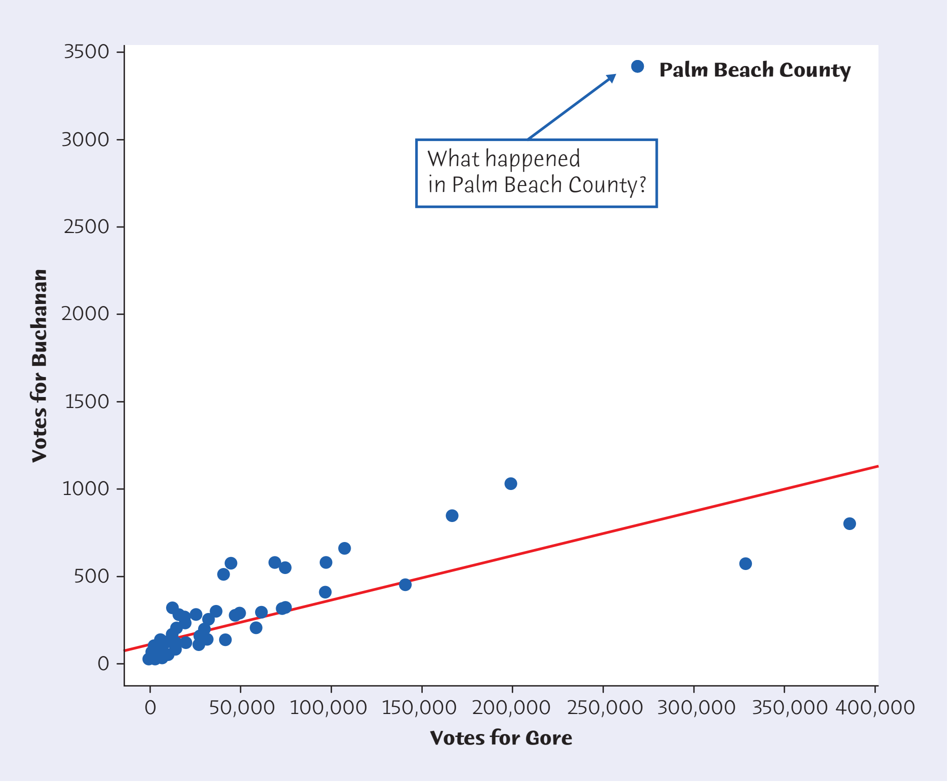
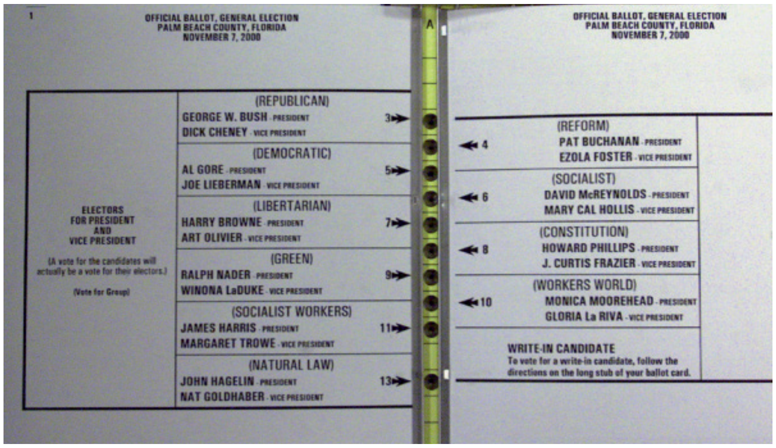


FIGURE 0.1 Votes in the 2000 presidential election for Al Gore and Patrick Buchanan in Florida's 67 counties. What happened in Palm Beach County?

What happened in Palm Beach County? The question leaps out from the graph. In this large and heavily Democratic county, a conservative third-party candidate did far better relative to the Democratic candidate than in any other county. The points for the other 66 counties show votes for both candidates increasing together in a roughly straight-line pattern. Both counts go up as county population goes up. Based on this pattern, we would expect Buchanan to receive around 800 votes in Palm Beach County. He actually received more than 3400 votes. That difference determined the election result in Florida and in the nation. ■

The graph demands an explanation. It turns out that Palm Beach County used a confusing "butterfly" ballot (see photo on page 5), in which candidate names on both left and right pages led to a voting column in the center. It would be easy for a voter who intended to vote for Gore to in fact cast a vote for Buchanan. The graph is convincing evidence that this in fact happened.

Most statistical software will draw a variety of graphs with a few simple commands. Examining your data with appropriate graphs and numerical summaries is the correct place to begin most data analyses. These can often reveal important patterns or trends that will help you understand what your data has to say.



Bruce Weaver/AFP/Getty Images

0.3 Variation is everywhere

The company's sales reps file into their monthly meeting. The sales manager rises. "Congratulations! Our sales were up 2% last month, so we're all drinking champagne this morning. You remember that when sales were down 1% last month I fired half of our reps." This picture is only slightly exaggerated. Many managers overreact to small short-term variations in key figures. Here is Arthur Nielsen, former head of the country's largest market research firm, describing his experience:

> *Too many business people assign equal validity to all numbers printed on paper. They accept numbers as representing Truth and find it difficult to work with the concept of probability. They do not see a number as a kind of shorthand for a range that describes our actual knowledge of the underlying condition.*[3]

Business data such as sales and prices vary from month to month for reasons ranging from the weather to a customer's financial difficulties to the inevitable errors in gathering the data. The manager's challenge is to say when there is a real pattern behind the variation. We'll see that statistics provides tools for understanding variation and for seeking patterns behind the screen of variation. Let's look at some more data.

EXAMPLE 0.4 The Price of Gas

Figure 0.2 plots the average price of a gallon of regular unleaded gasoline each week from September 1990 to June 2013.[4] There certainly is variation! But a close look shows a yearly pattern: gas prices go up during the summer driving season, then down as demand drops in the fall. On top of this regular pattern, we see the effects of international events. For example, prices rose when the 1990 Gulf War threatened oil supplies and dropped when the world economy turned down after the September 11, 2001, terrorist attacks in the United States. The years 2007 and 2008 brought the perfect storm: the ability to produce oil and refine gasoline was overwhelmed by high demand from China and the United States and continued turmoil in the oil-producing areas of the Middle East and Nigeria. Add in a rapid fall in the value of the dollar, and prices at the pump skyrocketed to more than $4 per gallon. In 2010 the Gulf oil spill also affected supply and hence prices. The data carry an important message: because the United States imports much of its oil, we can't control the price we pay for gasoline. ■

Tony Avelar/Bloomberg via Getty Images

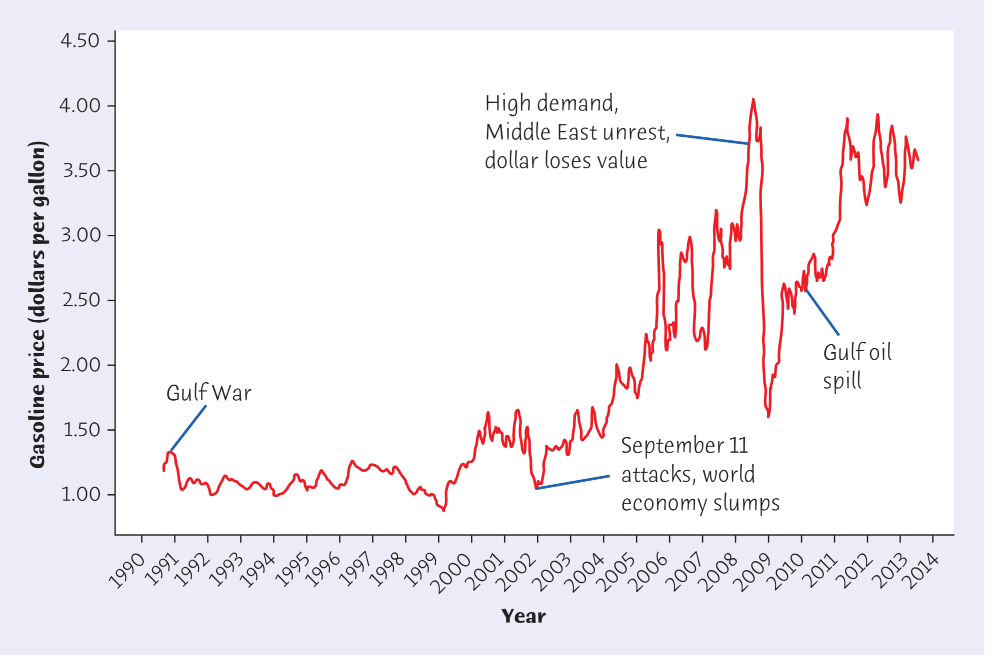


FIGURE 0.2
Variation is everywhere: the average retail price of regular unleaded gasoline, 1990 to mid 2013.

Variation is everywhere. Individuals vary; repeated measurements on the same individual vary; almost everything varies over time. One reason we need to know some statistics is that it helps us deal with variation and to describe the uncertainty in our conclusions. Let's look at another example to see how variation is incorporated into our conclusions.

EXAMPLE 0.5 The HPV Vaccine

Cervical cancer, once the leading cause of cancer deaths among women, is the easiest female cancer to prevent with regular screening tests and follow-up. Almost all cervical cancers are caused by human papillomavirus (HPV). The first vaccine to protect against the most common varieties of HPV became available in 2006. The Centers for Disease Control and Prevention recommend that all girls be vaccinated at age 11 or 12. In 2011, the CDC made the same recommendation for boys, to protect against anal and throat cancers caused by the HPV virus.

How well does the vaccine work? Doctors rely on experiments (called "clinical trials" in medicine) that give some women the new vaccine and others a dummy vaccine. (This is ethical when it is not yet known whether or not the vaccine is safe and effective.) The conclusion of the most important trial was that an estimated 98% of women up to age 26 who are vaccinated before they are infected with HPV will avoid cervical cancers over a three-year period.

Women who get the vaccine are much less likely to get cervical cancer. But because variation is everywhere, the results are different for different women. Some vaccinated women will get cancer, and many who are not vaccinated will escape. Statistical conclusions are "on the average" statements only, and even these "on the average" statements have an element of uncertainty. Although we can't be 100% certain that the vaccine reduces risk on the average, statistics allows us to state how confident we are that this is the case. ■

Because variation is everywhere, conclusions are uncertain. Statistics gives us a language for talking about uncertainty that is used and understood by statistically literate people everywhere. In the case of HPV vaccine, the medical journal used that language to tell us: “Vaccine efficiency . . . was 98% (95 percent confidence interval 86% to 100%).”[5] That “98% effective” is, in Arthur Nielsen’s words, “shorthand for a range that describes our actual knowledge of the underlying condition.” The range is 86% to 100%, and we are 95 percent confident that the truth lies in that range. We will soon learn to understand this language. We can’t escape variation and uncertainty. Learning statistics enables us to live more comfortably with these realities.

0.4 What lies ahead in this book

The purpose of *Basic Practice of Statistics* is to give you a working knowledge of the ideas and tools of practical statistics. We will divide practical statistics into three main areas.

- **Data analysis** concerns methods and strategies for looking at data; exploring, organizing, and describing data using graphs and numerical summaries. Your thoughtful exploration allows data to illuminate reality. Part I of this book (Chapters 1 to 6) discusses data analysis.
- **Data production** provides methods for producing data that can give clear answers to specific questions. Where data come from matters and is often the most important limitation on their usefulness. Basic concepts about how to select samples and design experiments are some of the most influential ideas in statistics. These concepts are the subject of Chapters 8 and 9.
- **Statistical inference** moves beyond the data in hand to draw conclusions about some wider universe. Statistical conclusions aren’t yes-or-no answers—they must take into account that variation is everywhere; variability among people, animals, or objects and uncertainty in data. To describe variation and uncertainty, inference uses the language of probability, introduced in Chapter 12. Because we are concerned with practice rather than theory, we need only a limited knowledge of probability. Chapters 13 and 14 offer more probability for those who want it. Chapters 15 to 18 discuss the reasoning of statistical inference. These chapters are the key to the rest of the book. Chapters 20 to 23 present inference as used in practice in the most common settings. Chapters 25 to 27 concern more advanced or specialized kinds of inference.

Because data are numbers with a context, doing statistics means more than manipulating numbers. You must **state** a problem in its real-world context, **plan** your specific statistical work in detail, **solve** the problem by making the necessary graphs and calculations, and **conclude** by explaining what your findings say about the real-world setting. We’ll make regular use of this four-step process to encourage good habits that go beyond graphs and calculations to ask, “What do the data tell me?”

Statistics does involve lots of calculating and graphing. The text presents the techniques you need, but you should use technology to automate calculations and graphs as much as possible. Because the big ideas of statistics don’t depend on any particular level of access to technology, *Basic Practice of Statistics* does not require software or a graphing calculator until we reach the more advanced methods in Part V of the text. Even if you make little use of technology, you should look at the “Using Technology” sections throughout the book. You will see at once that you can read and apply the output from almost any technology used for statistical calculations. The ideas really are more important than the details of how to do the calculations.

Unless you have access to software or a graphing calculator, *you will need a basic calculator with some built-in statistical functions.* Specifically, your calculator should find means and standard deviations and calculate correlations and regression lines. Look for a calculator that claims to do "two-variable statistics" or mentions "regression."

Although ability to carry out statistical procedures is very useful in academics and employment, the most important asset you can gain from the study of statistics is an understanding of the big ideas about working with data. *Basic Practice of Statistics* tries to explain the most important ideas of statistics, not just teach methods. Some examples of big ideas that you will meet (one from each of the three areas of statistics) are "always plot your data," "randomized comparative experiments," and "statistical significance."

You learn statistics by doing statistical problems. As you read, you will see several levels of exercises, arranged to help you learn. Short "Apply Your Knowledge" problem sets appear after each major idea. These are straightforward exercises that help you solidify the main points as you read. Be sure you can do these exercises before going on. The end-of-chapter exercises begin with multiple-choice "Check Your Skills" exercises (with all answers in the back of the book). Use them to check your grasp of the basics. The regular "Chapter Exercises" help you combine all the ideas of a chapter. Finally, the four Part Review chapters (Chapters 7, 11, 19, and 24) look back over major blocks of learning, with many review exercises. At each step you are given less advance knowledge of exactly what statistical ideas and skills the problems will require, so each type of exercise requires more understanding.

The key to learning is persistence. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. The gain will be worth the pain.