W. H. Freeman and Company
New York

The Basic Practice of Statistics

Fourth Edition

David S. Moore
Purdue University

Publisher: **Craig Bleyer**
Executive Editor: **Ruth Baruth**
Associate Acquisitions Editor: **Laura Hanrahan**
Marketing Manager: **Victoria Anderson**
Editorial Assistant: **Laura Capuano**
Photo Editor: **Bianca Moscatelli**
Photo Researcher: **Brian Donnelly**
Cover and Text Designer: **Vicki Tomaselli**
Cover and Interior Illustrations: **Mark Chickinelli**
Senior Project Editor: **Mary Louise Byrd**
Illustration Coordinator: **Bill Page**
Illustrations: **Techbooks**
Production Manager: **Julia DeRosa**
Composition: **Techbooks**
Printing and Binding: **Quebecor World**

Library of Congress Control Number: 2006926755

ISBN: 0-7167-7478-X (Hardcover)
EAN: 9780716774785 (Hardcover)

ISBN: 0-7167-7463-1 (Softcover)
EAN: 978-0-7167-7463-1 (Softcover)

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 optional.

Contents

*Starred material is optional.

To The Instructor: About This Book

The Basic Practice of Statistics (*BPS*) is an introduction to statistics for college and university students that emphasizes balanced content, working with real data, and statistical ideas. It is designed to be accessible to students with limited quantitative background—just "algebra" in the sense of being able to read and use simple equations. The book 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. *BPS* was the pioneer in presenting a modern approach to statistics in a genuinely elementary text. In the following I describe for instructors the nature and features of the book and the changes in this fourth edition.

Guiding principles

BPS is based on three principles: balanced content, experience with data, and the importance of ideas.

Balanced content. Once upon a time, basic statistics courses taught probability and inference almost exclusively, often preceded by just a week of histograms, means, and medians. Such unbalanced content does not match 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 also good pedagogical reasons for beginning with data analysis (Chapters 1 to 7), then moving to data production (Chapters 8 and 9), and then to probability (Chapters 10 to 13) and inference (Chapters 14 to 29). 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. *BPS* gives a full presentation of basic probability and inference (20 of the 29 chapters) but places it in the context of statistics as a whole.

Experience with 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. *BPS* 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 *BPS*, though intended for beginners, use real data and give enough background to allow students to consider the meaning of their calculations. Even the first examples carry a message: a look at Arbitron data on radio station formats (page 7) and on use of portable music players in several age groups (page 8) shows that the Arbitron data don't help plan advertising for a music-downloading Web site. 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. *BPS* begins to develop students' judgment about statistical studies.

The importance of ideas. 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 her college stat course. *BPS* 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.

These three principles are widely accepted by statisticians concerned about teaching. In fact, statisticians have reached a broad consensus that first courses should reflect how statistics is actually used. As Richard Scheaffer says in discussing a survey paper of mine, "With regard to the content of an introductory statistics course, statisticians are in closer agreement today than at any previous time in my career."[1*] Figure 1 is an outline of the consensus as summarized by the Joint Curriculum Committee of the American Statistical Association and the Mathematical Association of America.[2] I was a member of the ASA/MAA committee, and I agree with their conclusions. More recently, the College Report of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project has emphasized exactly the same themes.[3] Fostering active learning is the business of the teacher, though an emphasis on working with data helps. *BPS* is guided by the content emphases of the modern consensus. In the language of the GAISE recommendations, these are: develop statistical thinking, use real data, stress conceptual understanding.

Accessibility

The intent of *BPS* 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 them everything. Another principle is to offer frequent stopping points. *BPS* presents its content in relatively short chapters, each ending with a summary and two levels of exercises. Within chapters, a few "Apply Your Knowledge" exercises follow each new idea or skill for a quick check of basic

APPLY YOUR KNOWLEDGE

* All notes are collected in the Notes and Data Sources section at the end of the book.

1. **Emphasize the elements of statistical thinking:**
   (a) the need for data;
   (b) the importance of data production;
   (c) the omnipresence of variability;
   (d) the measuring and modeling of variability.
2. **Incorporate more data and concepts, fewer recipes and derivations. Wherever possible, automate computations and graphics**. An introductory course should:
   (a) rely heavily on *real* (not merely realistic) data;
   (b) emphasize *statistical* concepts, e.g., causation vs. association, experimental vs. observational, and longitudinal vs. cross-sectional studies;
   (c) rely on computers rather than computational recipes;
   (d) treat formal derivations as secondary in importance.
3. **Foster active learning,** through the following alternatives to lecturing:
   (a) group problem solving and discussion;
   (b) laboratory exercises;
   (c) demonstrations based on class-generated data;
   (d) written and oral presentations;
   (e) projects, either group or individual.

FIGURE 1 Recommendations of the ASA/MAA Joint Curriculum Committee.

mastery—and also to mark off digestible bites of material. Each of the first three parts of the book ends with a review chapter that includes a point-by-point outline of skills learned and many review exercises. (Instructors can choose to cover any or none of the chapters in Parts IV and V, so each of these chapters includes a skills outline.) The review chapters present many additional exercises without the "I just studied that" context, thus asking for another level of learning. I think it is helpful to assign some review exercises. Look at the first five exercises of Chapter 22 (the Part III review) to see the advantage of the part reviews. Many instructors will find that the review chapters appear at the right points for pre-examination review.

Technology

Automating calculations increases students' ability to complete problems, reduces their frustration, and helps them concentrate on ideas and problem recognition rather than mechanics. *All students should have at least a "two-variable statistics" calculator* with functions for correlation and the least-squares regression line as well as for the mean and standard deviation. Because students have calculators, the text doesn't discuss out-of-date "computing formulas" for the sample standard deviation or the least-squares regression line.

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. *BPS* does not assume or require use of software except in Chapters 24 and 25, 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 displays and comments on output from the same four technologies, representing graphing calculators (the Texas Instruments TI-83 or TI-84), spreadsheets (Microsoft Excel), and statistical software (CrunchIt! and Minitab). The output always concerns one of the main teaching examples, so that students can compare text and output.

Using technology

A quite different use of technology appears in the interactive applets created to my specifications and available online and on the text CD. These are designed primarily to help in learning statistics rather than in doing statistics. An icon calls attention to comments and exercises based on the applets. I suggest using selected applets for classroom demonstrations even if you do not ask students to work with them. The *Correlation and Regression*, *Confidence Interval*, and new *P-value* applets, for example, convey core ideas more clearly than any amount of chalk and talk.

APPLET

What's new?

BPS has been very successful. There are no major changes in the statistical content of this new edition, but longtime users will notice the following:

- **Many new examples and exercises.**
- **Careful rewriting** with an eye to yet greater clarity. Some sections, for example, Normal calculations in Chapter 3 and power in Chapter 16, have been completely rewritten.
- **A new commentary on Data Ethics** following Chapter 9. Students are increasingly aware that science often poses ethical issues. Instruction in science should therefore not ignore ethics. Statistical studies raise questions about privacy and protection of human subjects, for example. The commentary describes such issues, outlines accepted ethical standards, and presents striking examples for discussion.

In preparing this edition, I have concentrated on pedagogical enhancements designed to make it easier for students to learn.

- **A handy "Caution" icon** in the margin calls attention to common confusions or pitfalls in basic statistics.

CAUTION

- **Many small marginal photos** are chosen to enhance examples and exercises. Students see, for example, a water-monitoring station in the Everglades (page 22) or a *Heliconia* flower (page 54) when they work with data from these settings.

Check Your Skills

- **A set of "Check Your Skills" multiple-choice items** opens each set of chapter exercises. These are deliberately straightforward, and answers to all appear in the back of the book. Have your students use them to assess their grasp of basic ideas and skills, or employ them in a "clicker" classroom response system for class review.

4 STEP

- **A new four-step process** (State, Formulate, Solve, Conclude) guides student work on realistic statistical problems. See the inside front cover for an overview. I outline and illustrate the process early in the text (see page 53), but its full usefulness becomes clear only as we accumulate the tools needed for realistic problems. In later chapters this process organizes most examples and many exercises. The process emphasizes a major theme in *BPS:* statistical problems originate in a real-world setting ("State") and require conclusions in the language of that setting ("Conclude"). Translating the problem into the formal language of statistics ("Formulate") is a key to success. The graphs and computations needed ("Solve") are essential but not the whole story. A marginal icon helps students see the four-step process as a thread through the text. I have been careful not to let this outline stand in the way of clear exposition. Most examples and exercises, especially in earlier chapters, intend to teach specific ideas and skills for which the full process is not appropriate. It is absent from some entire chapters (for example, those on probability) where it is not relevant. Nonetheless, the cumulative effect of this overall strategy for problem solving should be substantial.

CrunchIt!

- **CrunchIt! statistical software** is available online with new copies of *BPS*. Developed by Webster West of Texas A&M University, CrunchIt! offers capabilities well beyond those needed for a first course. It implements modern procedures presented in *BPS*, including the "plus four" confidence intervals for proportions. More important, I find it the easiest true statistical software for student use. Check out, for example, CrunchIt!'s flexible and straightforward process for entering data, often a real barrier to software use. I encourage teachers who have avoided software in the past for reasons of availability, cost, or complexity to consider CrunchIt!.

Why did you do that?

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

Why does the distinction between population and sample not appear in Part I? This is a sign that 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. I see no need for a special word, but I think Tukey is right.

Why not begin with data production? It is certainly reasonable to do so—the natural flow of a planned study is from design to data analysis to inference. But in their future employment most students will use statistics mainly in settings other than planned research studies. I place the design of data production (Chapters 8 and 9) after data analysis to emphasize that data-analytic techniques apply to any data. One of the primary purposes of statistical designs for producing data is to make inference possible, so the discussion in Chapters 8 and 9 opens Part II and motivates the study of probability.

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. I prefer not to suggest that this material is essentially tied to probability, as the traditional order does. And I find it very 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 is traditional? *BPS* 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 a firm's employees. Perhaps most important, the *BPS* approach 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. I 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. I consider Chapters 4 and 5 (correlation and regression) essential and Chapter 24 (regression inference) optional.

What about probability? Much of the usual formal probability appears in the *optional* Chapters 12 and 13. Chapters 10 and 11 present in a less formal way the ideas of probability and sampling distributions that are needed to understand inference. These two chapters follow a straight line from the idea of probability as long-term regularity, through concrete ways of assigning probabilities, to the central idea of the sampling distribution of a statistic. The law of large numbers and the central limit theorem appear in the context of discussing the sampling distribution of a sample mean. What is left to Chapters 12 and 13 is mostly "general probability rules" (including conditional probability) and the binomial distributions.

I suggest that you omit Chapters 12 and 13 unless you are constrained by external forces. Experienced teachers recognize that students find probability difficult. Research on learning confirms our experience. Even students who can do formally posed probability problems often have a very fragile conceptual grasp of probability ideas. Attempting to present a substantial introduction to probability in a data-oriented statistics course for students who are not mathematically trained is in my opinion unwise. Formal probability does not help these students master the ideas of inference (at least not as much as we teachers often imagine), and it depletes reserves of mental energy that might better be applied to essentially statistical ideas.

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. Sometime in the golden future we will start with resampling methods. I think that permutation tests make the reasoning of tests clearer than any traditional approach. For now the main choices are z for a mean and z for a proportion.

I find z for means quite a bit more accessible to students. Positively, we can say up front that we are going to explore the reasoning of inference in an overly simple setting. Remember, 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 exactly Normal sampling distribution, but a Normal approximation to a discrete distribution; use of $\hat{p}$ in both the numerator and the 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. Once upon a time we had at least the compensation of developing practically useful procedures. Now the often gross inaccuracy of the traditional z confidence interval for p is better understood. See the following explanation.

Why does the presentation of inference for proportions go beyond the traditional methods? Recent computational and theoretical work has demonstrated convincingly that the standard confidence intervals for proportions can be trusted only for very large sample sizes. It is hard to abandon old friends, but I think that a look at the graphs in Section 2 of the paper by Brown, Cai, and DasGupta in the May 2001 issue of *Statistical Science* is both distressing and persuasive.[4] The standard intervals often have a true confidence level much less than what was requested, and requiring larger samples encounters a maze of "lucky" and "unlucky" sample sizes until very large samples are reached. Fortunately, there is a simple cure: just add two successes and two failures to your data. I present these "plus four intervals" in Chapters 20 and 21, along with guidelines for use.

Why didn't you cover Topic X? Introductory texts ought not to be encyclopedic. Including each reader's favorite special topic results in a text that is formidable in size and intimidating to students. I 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 *BPS*, Chapters 1 to 11 and 14 to 22, will have little difficulty moving on to more elaborate methods. There are of course seven additional chapters in *BPS*, three in this volume and four available on CD and/or online, to guide the next stages of learning.

I am grateful to the many colleagues from two-year and four-year colleges and universities who commented on successive drafts of the manuscript. Special thanks are due to Patti Collings (Brigham Young University), Brad Hartlaub (Kenyon College), and Dr. Jackie Miller (The Ohio State University), who read the manuscript line by line and offered detailed advice. Others who offered comments are:

Holly Ashton,
Pikes Peak Community College
Sanjib Basu,
Northern Illinois University
Diane L. Benner,
Harrisburg Area Community College
Jennifer Bergamo,
Cicero-North Syracuse High School
David Bernklau,
Long Island University, Brooklyn Campus
Grace C. Cascio-Houston, Ph.D.,
Louisiana State University at Eunice
Dr. Smiley Cheng,
University of Manitoba
James C. Curl,
Modesto Junior College
Nasser Dastrange,
Buena Vista University
Mary Ellen Davis,
Georgia Perimeter College
Dipak Dey,
University of Connecticut
Jim Dobbin,
Purdue University
Mark D. Ecker,
University of Northern Iowa
Chris Edwards,
University of Wisconsin, Oshkosh
Teklay Fessahaye,
University of Florida
Amy Fisher,
Miami University, Middletown
Michael R. Frey,
Bucknell University
Mark A. Gebert, Ph.D.,
Eastern Kentucky University
Jonathan M. Graham,
University of Montana
Betsy S. Greenberg,
University of Texas, Austin
Ryan Hafen,
University of Utah
Donnie Hallstone,
Green River Community College
James Higgins,
Kansas State University
Lajos Horvath,
University of Utah

Patricia B. Humphrey,
University of Alaska
Lloyd Jaisingh,
Morehead State University
A. Bathi Kasturiarachi,
Kent State University, Stark Campus
Mohammed Kazemi,
University of North Carolina, Charlotte
Justin Kubatko,
The Ohio State University
Linda Kurz,
State University of New York, Delhi
Michael Lichter,
University of Buffalo
Robin H. Lock,
St. Lawrence University
Scott MacDonald,
Tacoma Community College
Brian D. Macpherson,
University of Manitoba
Steve Marsden,
Glendale Community College
Kim McHale,
Heartland Community College
Kate McLaughlin,
University of Connecticut
Nancy Role Mendell,
State University of New York, Stonybrook
Henry Mesa,
Portland Community College
Dr. Panagis Moschopoulos,
The University of Texas, El Paso
Kathy Mowers,
Owensboro Community and Technical College
Perpetua Lynne Nielsen,
Brigham Young University
Helen Noble,
San Diego State University
Erik Packard,
Mesa State College
Christopher Parrett,
Winona State University
Eric Rayburn,
Danville Area Community College
Dr. Therese Shelton,
Southwestern University
Thomas H. Short,
Indiana University of Pennsylvania
Dr. Eugenia A. Skirta,
East Stroudsburg University
Jeffrey Stuart,
Pacific Lutheran University
Chris Swanson,
Ashland University
Mike Turegun,
Oklahoma City Community College
Ramin Vakilian,
California State University, Northridge
Kate Vance,
Hope College
Dr. Rocky Von Eye,
Dakota Wesleyan University
Joseph J. Walker,
Georgia State University

I am particularly grateful to Craig Bleyer, Laura Hanrahan, Ruth Baruth, Mary Louise Byrd, Vicki Tomaselli, Pam Bruton, and the other editorial and design professionals who have contributed greatly to the attractiveness of this book.

Finally, I am indebted to the many statistics teachers with whom I have discussed the teaching of our subject over many years; to people from diverse fields with whom I have worked to understand data; and especially to students whose compliments and complaints have changed and improved my teaching. Working with teachers, colleagues in other disciplines, and students constantly reminds me 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

Media and Supplements

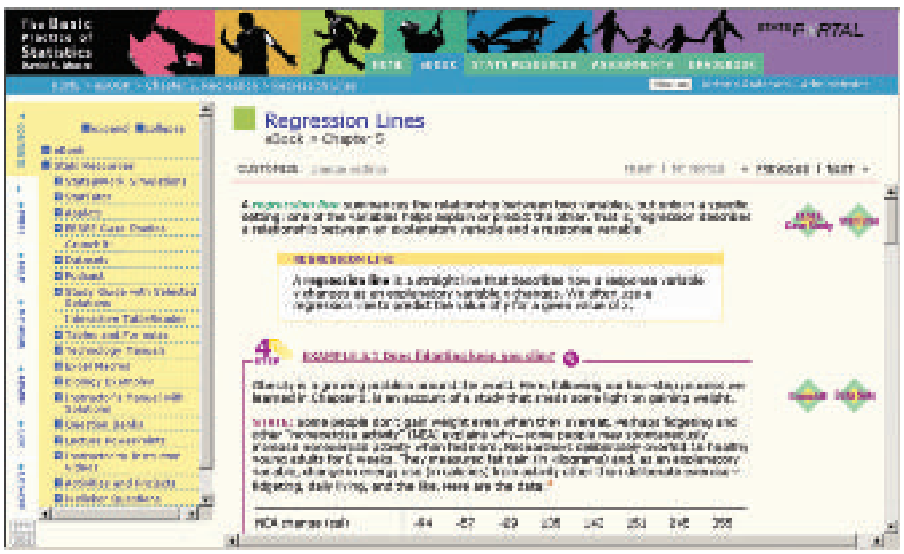
For students

A full range of media and supplements is available to help students get the most out of *BPS*. Please contact your W. H. Freeman representative for ISBNs and value packages.

NEW!

STATS PORTAL

One click. One place. For all the statistical tools you need.



www.whfreeman.com/statsportal (Access code required. Available packaged with *The Basic Practice of Statistics 4th Edition* or for purchase online.)

StatsPortal is the digital gateway to *BPS 4e*, designed to enrich your course and enhance your students' study skills through a collection of Web-based tools. StatsPortal integrates a rich suite of diagnostic, assessment, tutorial, and enrichment features, enabling students to master statistics at their own pace. Organized around three main teaching and learning components:

- **Interactive eBook** offers a complete online version of the text, fully integrated with all of the media resources available with *BPS 4e*.

- **StatsResource Center** organizes all of the resources for *BPS 4e* into one location for the student's ease of use. Includes:
  - **Stats@Work Simulations** put the student in the role of the statistical consultant, helping them better understand statistics interactively within the context of real-life scenarios. Students will be asked to interpret and analyze data presented to them in report form, as well as to interpret current event news stories. All tutorials are graded and offer helpful hints and feedback.
  - **StatTutor Tutorials** offer 84 audio-embedded tutorials tied directly to the textbook, containing videos, applets, and animations.
  - **Statistical Applets** these sixteen interactive applets help students master statistics interactively.
  - **EESEE Case Studies** developed by The Ohio State University Statistics Department provide students with a wide variety of timely, real examples with real data. Each case study is built around several thought-provoking questions that make students think carefully about the statistical issues raised by the stories.
  - **Podcast Chapter Summary** provides students with an audio version of chapter summaries so they can download and review on their mp3 player!
  - **CrunchIt! Statistical Software** allows users to analyze data from any Internet location. Designed with the novice user in mind, the software is not only easily accessible but also easy to use. Offers all the basic statistical routines covered in the introductory statistics courses and more!
  - **Datasets** are offered in ASCII, Excel, JMP, Minitab, TI, SPSS, S-Plus, Minitab, ASCII, and Excel format.
  - **Online Tutoring with SmarThinking** is available for homework help from specially trained, professional educators.
  - **Student Study Guide with Selected Solutions** includes explanations of crucial concepts and detailed solutions to key text problems with step-through models of important statistical techniques.
  - **Statistical Software Manuals** for TI-83, Minitab, Excel, and SPSS provide chapter-to-chapter applications and exercises using specific statistical software packages with *BPS 4e*.
  - **Interactive Table Reader** allows students to use statistical tables interactively to seek the information they need.
  - **Tables and Formulas** provide each table and formulas from the chapter.
  - **Excel Macros.**

StatsResources (instructor-only)

- **Instructor's Manual with Full Solutions** includes worked-out solutions to all exercises, teaching suggestions, and chapter comments.

    - **Test Bank** contains complete solutions for textbook exercises.
    - **Lecture PowerPoint Slides** gives instructors detailed slides to use in lectures.
    - **Activities and Projects** offers ideas for projects for Web-based exploration asking students to write critically about statistics.
    - **i>clicker Questions** these conceptually-based questions help instructors to query students using i>clicker's personal response units in class lectures.
    - **Instructor-to-Instructor Videos** provide instructors with guidance on how to use these interactive examples in the classroom.
    - **Biology Examples** identify areas of *BPS 4e* that relate to the field of biology.
- **Assignment Center** organizes assignments and guides instructors through an easy-to-create assignment process providing access to questions from the Test Bank, Check Your Skills, Apply Your Knowledge, Web Quizzes, and Exercises from *BPS 4e*. Enables instructors to create their own assignments from a variety of question-types for self-graded assignments. This powerful assignment manager allows instructors to select their preferred policies in regard to scheduling, maximum attempts, time limitations, feedback, and more!

New! Online Study Center: www.whfreeman.com/bps4e/osc (Access code required. Available for purchase online.) In addition to all the offerings available on the Companion Web site, the OSC offers:

- **StatTutor Tutorials**
- **CrunchIt! Statistical Software**
- **Stats@Work Simulations**
- **Study Guide**
- **Statistical Software Manuals**

The Companion Web Site: www.whfreeman.com/bps. Seamlessly integrates topics from the text. On this open-access Web site, students can find:

- **Interactive statistical applets** that allow students to manipulate data and see the corresponding results graphically.
- **Datasets** in ASCII, Excel, JMP, Minitab, TI, SPSS, and S-Plus formats.
- **Interactive exercises and self-quizzes** to help students prepare for tests.
- **Key tables and formulas** summary sheet.
- **All tables** from the text in .pdf format for quick, easy reference.

- **Additional exercises** for every chapter written by David Moore, giving students more opportunities to make sure they understand key concepts. Solutions to odd-numbered additional exercises are also included.
- **Optional Companion Chapters 26, 27, 28, and 29,** covering nonparametric tests, statistical process control, multiple regression, and two-way analysis of variance, respectively.
- **CrunchIt!** statistical software is available via an access-code-protected Web site. Access codes are available in every new text or can be purchased online for $5.
- **EESEE** case studies are available via an access-code-protected Web site. Access codes are available in every new text or can be purchased online.

Interactive Student CD-ROM: Included with every new copy of *BPS*, the CD contains access to most of the content available on the Web site. CrunchIt! statistical software and EESEE case studies are available via an access-code-protected Web site. (Access code is included with every new text.)

Special Software Packages: Student versions of JMP, Minitab, S-PLUS, and SPSS are available on a CD-ROM packaged with the textbook. This software is not sold separately and must be packaged with a text or a manual. Contact your W. H. Freeman representative for information or visit **www.whfreeman.com**.

NEW! SMARTHINKING Online Tutoring: (Access code required) W. H. Freeman and Company is partnering with SMARTHINKING to provide students with free online tutoring and homework help from specially trained, professional educators. Twelve-month subscriptions are available to be packaged with BPS.

The following supplements are available in print:

- **Student Study Guide** with Selected Solutions.
- **Activities and Projects Book.**

For instructors

The **Instructor's Web site** requires user registration as an instructor and features all of the student Web material plus:

- Instructor version of **EESEE** (Electronic Encyclopedia of Statistical Examples and Exercises), with solutions to the exercises in the student version.
- The **Instructor's Guide**, including full solutions to all exercises in .pdf format.
- **Text art images** in jpg format.

- **PowerPoint slides** containing textbook art embedded into each slide.
- **Lecture PowerPoint slides** offering a detailed lecture presentation of statistical concepts covered in each chapter of *BPS*.
- **Class Teaching Examples**, one or more new examples for each chapter of *BPS* with suggestions for classroom use by David Moore. Tables and graphs are in a form suitable for making transparencies.
- **Full solutions** to the more than 400 extra exercises in the **Additional Exercises** supplement on the student Web site.

Enhanced Instructor's Resource CD-ROM: Designed to help instructors create lecture presentations, Web sites, and other resources, this CD allows instructors to **search** and **export** all the resources contained below by key term or chapter:

- All text images
- Statistical applets, datasets, and more
- Instructor's Manual with full solutions
- PowerPoint files and lecture slides
- Test bank files

Annotated Instructor's Edition

Printed Instructor's Guide with Full Solutions

Test Bank: Printed or computerized (Windows and Mac on one CD-ROM).

Course Management Systems: W. H. Freeman and Company provides courses for Blackboard, WebCT (Campus Edition and Vista), and Angel course management systems. These are completely integrated solutions that you can easily customize and adapt to meet your teaching goals and course objectives. Upon request, we also provide courses for users of Desire2Learn and Moodle. Visit www.bfwpub.com/lms for more information.

NEW! i-clicker Radio Frequency Classroom Response System: Offered by W. H. Freeman and Company, in partnership with i-clicker, and created by educators for educators, i-clicker's system is the hassle-free way to make class time more interactive. Visit www.iclicker.com for more information.

Applications

The Basic Practice of Statistics presents a wide variety of applications from diverse disciplines. The list below indicates the number of examples and exercises which relate to different fields:

Examples

Agriculture: 8
Biological and environmental sciences: 25
Business and economics: 10
Education: 29
Entertainment: 5
People and places: 20
Physical sciences: 5
Political Science and public policy: 3
Psychology and behavioral sciences: 6
Public health and medicine: 33
Sports: 7
Technology: 16
Transportation and automobiles: 14

Exercises

Agriculture: 56
Biological and environmental sciences: 128
Business and economics: 145
Education: 162
Entertainment: 33
People and places: 168
Physical sciences: 23
Political Science and public policy: 37
Psychology and behavioral sciences: 22
Public health and medicine: 189
Sports: 36
Technology: 37
Transportation and automobiles: 65

For a complete index of applications of examples and exercises, please see the Annotated Instructor's Edition or the Web site: **www.whfreeman.com/bps**.

To the Student: Statistical Thinking


Stockbyte/PictureQuest

Statistics is about data. Data are numbers, but they are not "just numbers." **Data are numbers with a context.** The number 10.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 10.5 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 10.5 pounds is quite large, and that a human baby is unlikely to weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative.

Statistics is the science of data. 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 principles of statistical thinking.

DATA BEAT ANECDOTES

An anecdote is a striking story that sticks in our minds exactly because it is striking. Anecdotes humanize an issue, but they can be misleading.

Does living near power lines cause leukemia in children? The National Cancer Institute spent 5 years and $5 million gathering data on this question. The researchers compared 638 children who had leukemia with 620 who did not. They went into the homes and measured the magnetic fields in the children's bedrooms, in other rooms, and at the front door. They recorded facts about power lines near the family home and also near the mother's residence when she was pregnant. Result: no connection between leukemia and exposure to magnetic fields of the kind produced by power lines. The editorial that accompanied the study report in the *New England Journal of Medicine* thundered, "It is time to stop wasting our research resources" on the question.[1]

Now compare the effectiveness of a television news report of a 5-year, $5 million investigation against a televised interview with an articulate mother whose child has leukemia and who happens to live near a power line. In the public mind, the anecdote wins every time. A statistically literate person knows better. **Data are more reliable than anecdotes because they systematically describe an overall picture rather than focus on a few incidents.**

ALWAYS LOOK AT THE DATA

Yogi Berra 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.

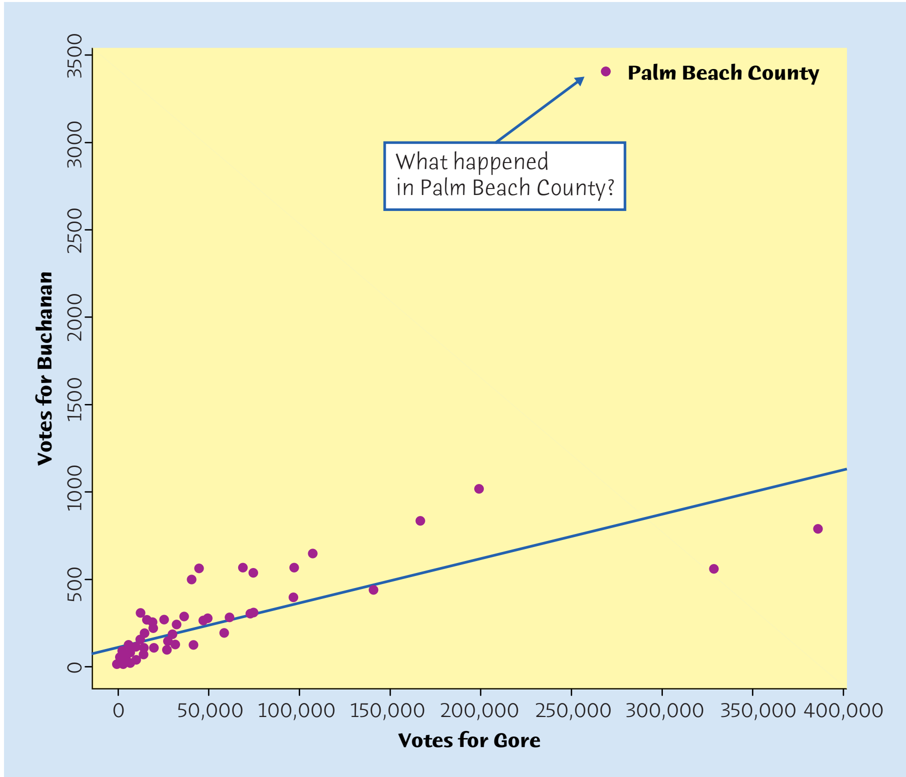


FIGURE 1 Votes in the 2000 presidential election for Al Gore and Patrick Buchanan in Florida's 67 counties. What happened in Palm Beach County?

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 election and made George Bush, rather than Al Gore, president. Let's look at some data. Figure 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.

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, 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, more convincing than the complaints of voters who (later) were unsure where their votes ended up.

BEWARE THE LURKING VARIABLE

The Kalamazoo (Michigan) Symphony once advertised a "Mozart for Minors" program with this statement: "Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music."[2] *Who would dispute that early experience with music builds brainpower?* The skeptical statistician, that's who. Children who take music lessons and attend concerts tend to have prosperous and well-educated parents. These same children are also likely to attend good schools, get good health care, and be encouraged to study hard. No wonder they score well on tests.

Brendan Byrne/Agefotostock

We call family background a *lurking variable* when we talk about the relationship between music and test scores. It is lurking behind the scenes, unmentioned in the symphony's publicity. Yet family background, more than anything else we can measure, influences children's academic performance. Perhaps the Kalamazoo Youth Soccer League should advertise that students who play soccer score higher on tests. After all, children who play soccer, like those who have experience in music, tend to have educated and prosperous parents. **Almost all relationships between two variables are influenced by other variables lurking in the background.**

WHERE THE DATA COME FROM IS IMPORTANT

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. *Do you believe that 70% of all parents regret having children?*

You shouldn't. The people who took the trouble to write Ann Landers are not representative of all parents. Their letters showed that many of them were angry at their children. All we know from these data is that there are some unhappy parents out there. A statistically designed poll, unlike Ann Landers's appeal, targets specific people chosen in a way that gives all parents the same chance to be asked. 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."

Here's another question: *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.

The evidence in favor of hormone replacement came from a number of studies that compared women who were taking hormones with others who were not. Beware the lurking variable: women who choose to take hormones are richer and better educated and see doctors more often than women who do not. These women do many things to maintain their health. It isn't surprising that they have fewer heart attacks.

To get convincing data on the link between hormone replacement and heart attacks, do an *experiment*. Experiments don't let women decide what to do. They assign women to either 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. By 2002, several experiments with women of different ages agreed that hormone replacement does *not* reduce the risk of heart attacks. The National Institutes of Health, after reviewing the evidence, concluded that the first studies were wrong. Taking hormones after menopause quickly fell out of favor.[3]

The most important information about any statistical study is how the data were produced. Only statistically designed opinion polls can be trusted. Only experiments can completely defeat the lurking variable and give convincing evidence that an alleged cause really does account for an observed effect.

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, 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.*[4]

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. Start by looking at the data.

Figure 2 plots the average price of a gallon of regular unleaded gasoline each month from January 1990 to February 2006.[5] There certainly is variation! But a close look shows a pattern: gas prices normally go up during the summer driving season each year, then down as demand drops in the fall. Against this regular pattern we see the effects of international events: prices rose because of the 1990 Gulf War and dropped because of the 1998 financial crisis in Asia and the September 11, 2001, terrorist attacks in the United States. The year 2005 brought the

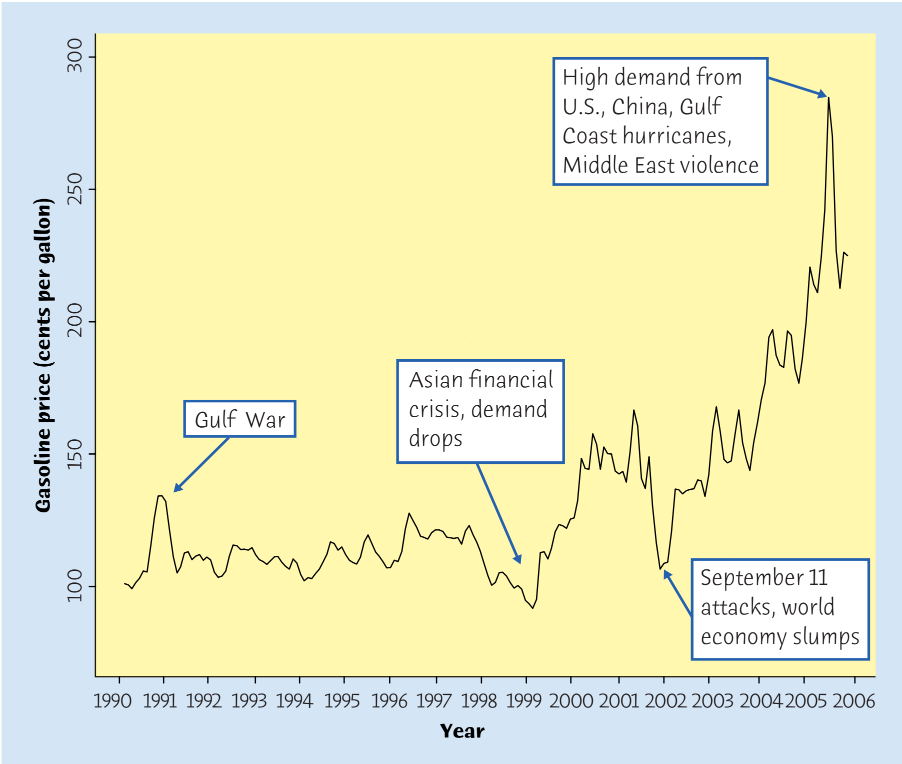
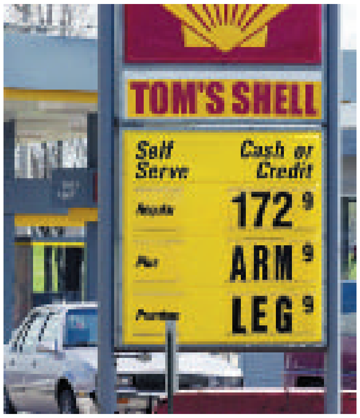


FIGURE 2 Variation is everywhere: the average retail price of regular unleaded gasoline, 1990 to early 2006.

perfect storm: the ability to produce oil and refine gasoline was overwhelmed by high demand from China and the United States, continued violence in Iraq, and hurricanes on the U.S. Gulf Coast. 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.

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 statistics helps us deal with variation.



AP/Wide World Photos

CONCLUSIONS ARE NOT CERTAIN

Most women who reach middle age have regular mammograms to detect breast cancer. *Do mammograms reduce the risk of dying of breast cancer?* To defeat the lurking variable, doctors rely on experiments (called "clinical trials" in medicine) that compare different ways of screening for breast cancer. The conclusion from 13 such trials is that mammograms reduce the risk of death in women aged 50 to 64 years by 26%.[6]

On the average, then, women who have regular mammograms are less likely to die of breast cancer. But because variation is everywhere, the results are different for different women. Some women who have yearly mammograms die of breast cancer, and some who never have mammograms live to 100 and die when they crash their motorcycles. Statistical conclusions are "on-the-average" statements only. Well then, can we be certain that mammograms reduce risk on the average? No. We can be very confident, but we can't be certain.

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 mammograms, the doctors use that language to tell us that "mammography reduces the risk of dying of breast cancer by 26 percent (95 percent confidence interval, 17 to 34 percent)." That 26% is, in Arthur Nielsen's words, a "shorthand for a range that describes our actual knowledge of the underlying condition." The range is 17% to 34%, 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.

Statistical Thinking and You

What Lies Ahead in This Book The purpose of *The Basic Practice of Statistics* (*BPS*) is to give you a working knowledge of the ideas and tools of practical statistics. We will divide practical statistics into three main areas:

1. **Data analysis** concerns methods and strategies for exploring, organizing, and describing data using graphs and numerical summaries. Only organized data can illuminate reality. Only thoughtful exploration of data can defeat the lurking variable. Part I of *BPS* (Chapters 1 to 7) discusses data analysis.
2. **Data production** provides methods for producing data that can give clear answers to specific questions. Where the data come from really is important. Basic concepts about how to select samples and design experiments are the most influential ideas in statistics. These concepts are the subject of Chapters 8 and 9.
3. **Statistical inference** moves beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and that conclusions are uncertain. To describe variation and uncertainty, inference uses the language of probability, introduced in Chapters 10 and 11. Because we are concerned with practice rather than theory, we need only a limited knowledge of probability. Chapters 12 and 13 offer more probability for those who want it. Chapters 14 to 16 discuss the reasoning of statistical inference. These chapters are the key to the rest of the book. Chapters 18 to 22 present inference as used in practice in the most common settings. Chapters 23 to 25, and the Optional Companion Chapters 26 to 29 on the text CD or online, 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, **formulate** the problem by recognizing what specific statistical work is needed, **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 a calculator or software 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 computing, *BPS* does not require software. 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 use 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.

You will need a 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."

Because graphing and calculating are automated in statistical practice, the most important assets you can gain from the study of statistics are an understanding of the big ideas and the beginnings of good judgment in working with data. *BPS* 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 three part review chapters 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 part review chapters (and the individual chapters in Part IV) include point-by-point lists of specific things you should be able to do. Go through that list, and be sure you can say "I can do that" to each item. Then try some of the review exercises. The book ends with a review titled "Statistical Thinking Revisited," which you should read and think about no matter where in the book your course ends.

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.