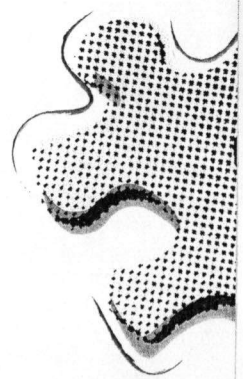
FIFTH EDITION

The Basic Practice of Statistics

DAVID S. MOORE

Purdue University

5th edition, 2010



W. H. Freeman and Company

New York

Course Management Systems

W. H. Freeman and Company provides course cartridges for Blackboard, WebCT (Campus Edition and Vista), and Angel course management systems. Upon request, we also provide courses for users of Desire2Learn and Moodle.

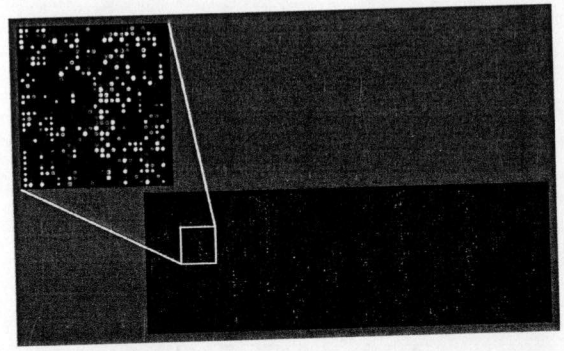
i>clicker Radio Frequency Classroom Response System
www.iclicker.com

Developed for educators by educators, i>clicker is the easiest-to-use and most flexible classroom response system available.

TO THE STUDENT

STATISTICAL THINKING

What genes are active in a tissue? Answering this question can unravel basic questions in biology, distinguish cancer cells from normal cells, and distinguish between closely related types of cancer. To learn the answer, apply the tissue to a "microarray" that contains thousands of snippets of DNA arranged in a grid on a chip about the size of your thumb. As DNA in the tissue binds to the snippets in the array, special recorders pick up spots of light of varying color and intensity across the grid and store what they see as numbers.



Paphrag at en.wikipedia

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 Web sites. When you buy a CD or download a digital track, the checkout scanner or Web site 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?

DNA microarrays, SoundScan, and medical studies all produce data (numerical facts), and lots of them. Using data effectively is a large and growing part of

most professions. Reacting to data is part of everyday life. That's why statistics is important:

STATISTICS IS THE SCIENCE OF LEARNING FROM 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.

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.[1]

WHERE THE DATA COME FROM MATTERS

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 are easier to arrange than careful *experiments*. 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. Part of the difficulty of a good experiment is persuading women to agree 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]

Of course, observational studies are often 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 rate 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. 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."

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.

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 (see page xxvi) 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 (see the illustration on page xxvi). 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.

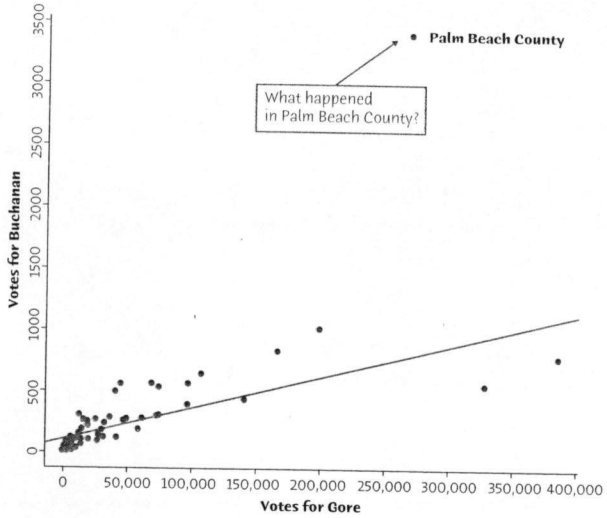
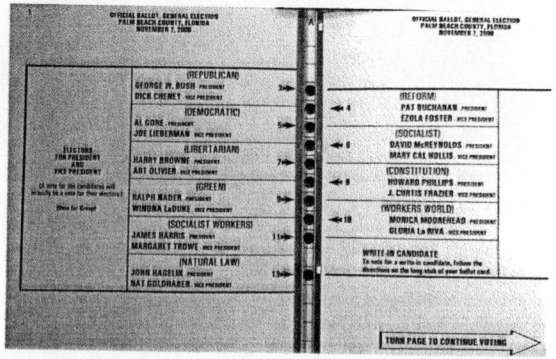


FIGURE 1

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



© Reuters/Corbis

BEWARE THE LURKING VARIABLE

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. Children who play soccer tend to have prosperous and well-educated parents. No wonder they do better in school (on the average) than children who don't play soccer. We can't conclude that hormone replacement reduces heart attacks or that playing soccer increases school grades just because we see these relationships in data. In both examples, education and affluence are *lurking variables*, background factors that help explain the relationships between hormone replacement and good health and between soccer and good grades.

Almost all relationships between two 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.

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.*[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. Figure 2 (see page xxviii) plots the average price of a gallon of regular unleaded gasoline each month from January 1990 to July 2008.[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

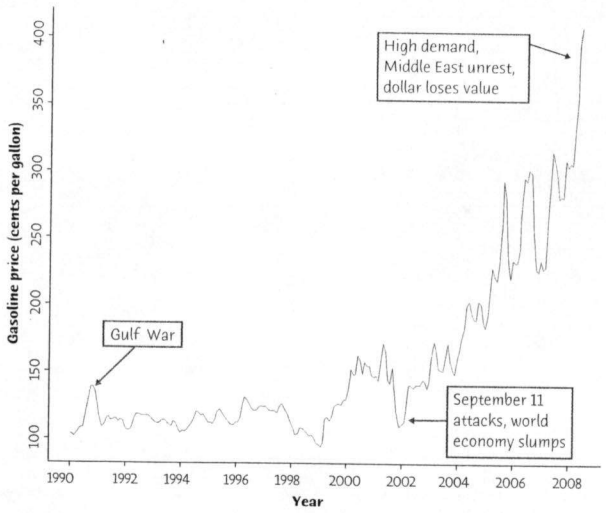


FIGURE 2
Variation is everywhere: the average retail price of regular unleaded gasoline, 1990 to early 2008.

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. The data carry an important message: because the United States imports most 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.

CONCLUSIONS ARE NOT CERTAIN

Cervical cancer is second only to breast cancer as a cause of cancer deaths in women. 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.

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 3-year period.

On the average 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. Well then, can we be certain that the vaccine reduces 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 HPV vaccine, the medical journal used that language to tell us that "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.

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 and 15 discuss the reasoning of statistical inference. These chapters are the key to the rest of the book. Chapters 17 to 20 present inference as used in practice in the most common settings. Chapters 22 to 24, and the Optional Companion Chapters 25 to 28 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, **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, *BPS* does not require software or a graphing calculator until we reach the more advanced methods in Part IV 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 constant 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-variables 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 (Chapters 7, 16, and 21) 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 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.