STATISTICAL THINKING

HOW TO TELL THE FACTS FROM THE ARTIFACTS

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 can't weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative.

Statistics uses data to gain insight and to draw conclusions. Our tools are graphs and calculations, but the tools 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 illuminate

What percent of the American population do you think is black? What percent do you think is white? When white Americans were asked these questions, their average answers were 23.8% black and 49.9% white. In fact, the Census Bureau tells us that 12.3% of Americans are black and 75.1% are white.[1]

Race remains a central social issue in the United States. It is illuminating to see that whites think (wrongly) that they are a minority. The census data—what really is true about the U.S. population—are also illuminating. We wonder if knowing the facts might help change attitudes.

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 the question. 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.[2]

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, however, knows that data are more reliable than anecdotes because they systematically describe an overall picture rather than focusing on a few incidents.

Beware the lurking variable

Air travelers would like their flights to arrive on time. Airlines collect data about on-time arrivals and report them to the Department of Transportation. Here are one month's data for flights from several western cities for two airlines:

	On time	Delayed
Alaska Airlines	3274	501
America West	6438	787

You can see that the percentages of late flights were:

$$\text{Alaska Airlines} \quad \frac{501}{3775} = 13.3\%$$

$$\text{America West} \quad \frac{787}{7225} = 10.9\%$$

It appears that America West does better.

This isn't the whole story, however. Almost all relationships between two variables are influenced by other variables lurking in the background. We have data on two variables, the airline and whether or not the flight was late. Let's add data on a third variable, which city the flight left from.[3]

	Alaska Airlines		America West	
	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61
Total	3274	501	6438	787

The "Total" row shows that the new table describes the same flights as the earlier table. Look again at the percentages of late flights, first for Los Angeles:

$$\text{Alaska Airlines} \quad \frac{62}{559} = 11.1\%$$

$$\text{AmericaWest} \quad \frac{117}{811} = 14.4\%$$

Alaska Airlines wins. The percentage delayed for Phoenix are:

$$\text{Alaska Airlines} \quad \frac{12}{233} = 5.2\%$$

$$\text{AmericaWest} \quad \frac{415}{5255} = 7.9\%$$

Alaska Airlines wins again. In fact, as Figure 1 shows, Alaska Airlines has a lower percentage of late flights at *every one* of these cities.

How can it happen that Alaska Airlines wins at every city but America West wins when we combine all the cities? Look at the data: America West flies most often from sunny Phoenix, where there are few delays. Alaska Airlines flies most often from Seattle, where fog and rain cause frequent delays. What city we fly from has a major influence on the chance of a delay, so including the city data reverses our conclusion. The message is worth repeating: almost all relationships between two variables are influenced by other variables lurking in the background.

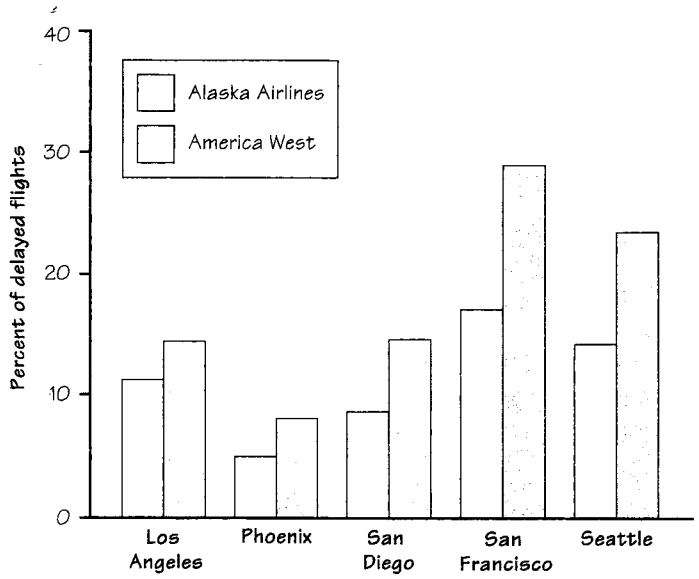

FIGURE 1 Comparing the percents of delayed flights for two airlines at five airports.

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.

The lesson: if you are careless about how you get your data, you may announce 70% "No" when the truth is close to 90% "Yes."

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. Statistical tools can help. Sometimes it is enough to simply plot the data. Figure 2 plots the average price of a gallon of unleaded gasoline each month for a decade. Behind the month-to-month variation you can see a gradual upward trend, the higher prices during each summer driving season, the upward spike during the 1990 Gulf War, and the sudden drop in the spring of 1998.[5]

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.

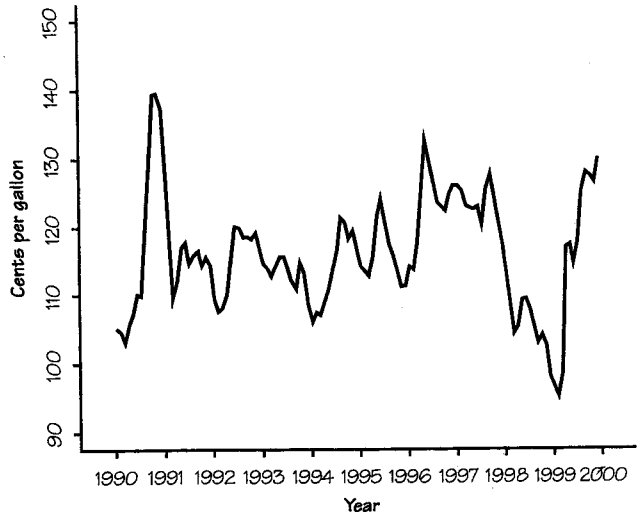

FIGURE 2 Variation is everywhere: the price at the pump of unleaded gasoline, January 1990 to December 1999.

Conclusions are not certain

Most women who reach middle age have regular mammograms to detect breast cancer. Do mammograms really reduce the risk of dying of breast cancer? To seek answers, doctors rely on "randomized clinical trials" that compare different ways of screening for breast cancer. We will see later that data from randomized comparative experiments are as good as it gets. 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 t die of breast cancer. But because variation is everywhere, the results are different fo different women. Some women who have mammograms every year die of breas cancer, and some who never have mammograms live to 100 and die when the crash their motorcycles. Can we be sure that mammograms reduce risk on the aver age? No, we can't be sure. *Because variation is everywhere, conclusions are uncertain*

Statistics gives us a language for talking about uncertainty that is used an understood by statistically literate people everywhere. In the case of mammograms the doctors use that language to tell us that "mammography reduces the risk o dying of breast cancer by 26 percent (95 percent confidence interval, 17 to 34 per cent)." That 26% is, in Arthur Nielsen's words, "a shorthand for a range tha describes our actual knowledge of the underlying condition." The range is 17% t 34%, and we are 95 percent confident that the truth lies in that range. We will soor 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 this book is to give you a working knowledge of the ideas and tools of practical statistics. We will divide practical statistics into four parts that reflect our short introduction to statistical thinking:

1. **Data analysis** concerns methods and strategies for exploring, organizing, and describing data using graphs and numerical summaries. Only organized data can illuminate. Only thoughtful exploration of data can defeat the lurking variable. Chapters 1 through 4 discuss 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 Chapter 5.

3. **Probability** is the language we use to describe chance, variation, and risk. Because variation is everywhere, probabilistic thinking helps separate reality from background noise. Chapters 6 through 9 present the essential ideas of probability.

4. **Statistical inference** moves beyond the data in hand to draw conclusions about a wider universe, taking into account that variation is everywhere and that conclusions are uncertain. Chapter 10 discusses the reasoning of statistical inference, and Chapters 11 through 15 present inference as used in practice in several settings.

Because data are numbers with a context, doing statistics means more than manipulating numbers. *The Practice of Statistics* is full of data, and each set of data has some brief background to help you understand what the data say. Examples and exercises usually express briefly some understanding gained from the data. In practice, you would know much more about the background

of the data you work with and about the questions you hope the data will answer. No textbook can be fully realistic. But it is important to form the habit of asking "What do the data tell me?" rather than just concentrating on making graphs and doing calculations. This book tries to encourage good habits.

Nonetheless, statistics involves lots of calculating and graphing. The text presents the techniques you need, but you should use a calculator or computer software to automate calculations and graphs as much as possible.

Ideas and judgment can't (at least yet) be automated. They guide you in telling the computer what to do and in interpreting its output. This book tries to explain the most important ideas of statistics, not just teach methods.

You learn statistics by doing statistical problems. This book offers three types of exercises, arranged to help you learn. Short problem sets appear after each major idea. These are straightforward exercises that help you solidify the main points before going on. The Section Exercises at the end of each numbered section help you combine all the ideas of the section. Finally, the Chapter Review Exercises look back over the entire chapter. At each step you are given less advance knowledge of exactly what statistical ideas and skills the problems will require, so each step requires more understanding. Each chapter ends with a Summary section that includes a detailed list of specific things you should now be able to do. Go through that list, and be sure you can say "I can do that" to each item. Then try some chapter exercises.

The basic principle of 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.

NOTES AND DATA SOURCES

PREFACE NOTES

1. D. S. Moore and discussants, "New pedagogy and new content: the case of statistics" *International Statistics Review*, 65 (1997), pp. 123–165. Richard Scheaffer's comment appears on page 156.
2. This summary of the committee's report was unanimously endorsed by the Board of Directors of the American Statistical Association. The full report is George Cobb, "Teaching statistics," in L. A. Steen (ed.), *Heeding the Call for Change: Suggestions for Curricular Action*, Mathematical Association of America, Washington, D.C., 1990, pp. 3–43.
3. David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 3rd ed., W. H. Freeman, New York, 1999.
4. A. C. Nielsen, Jr., "Statistics in Marketing," in *Making Statistics More Effective in Schools of Business*, Graduate School of Business, University of Chicago, 1986.
5. The data in Figure 2 are based on a component of the Consumer Price Index, from the Bureau of Labor Statistics on-line data center at http://stats.bls.gov:80/datahome.htm. To convert the index numbers into dollars, I used the national average price of $1.131 per gallon for January 1998, from the Energy Information Agency Web site, http://www.eia.doe.gov/.
6. H. C. Sox, "Editorial: benefit and harm associated with screening for breast cancer," *New England Journal of Medicine*, 338, No. 16 (1998).