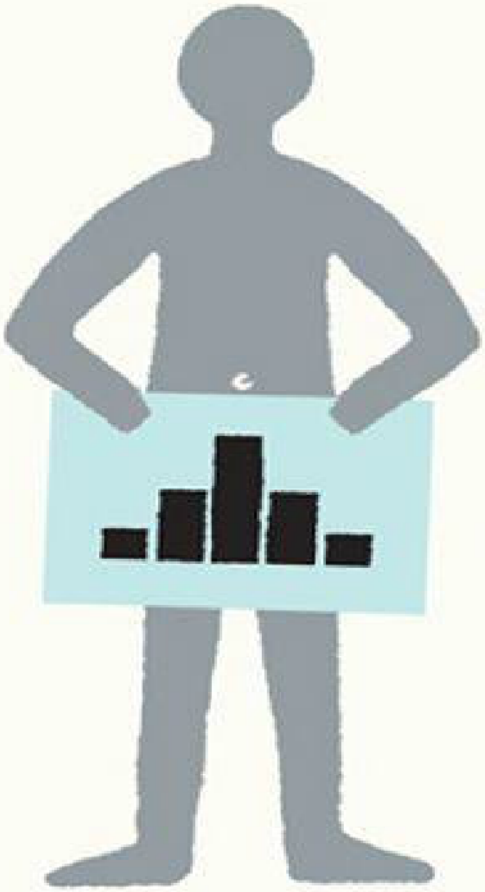
naked statistics

STRIPPING THE DREAD FROM THE DATA



charles wheelan

BEST-SELLING AUTHOR OF *NAKED ECONOMICS*

naked statistics

Stripping the Dread from the Data

CHARLES WHEELAN

W. W. Norton & Company
New York | London

Dedication

For Katrina

Contents

Cover
Title Page
Dedication

Introduction: *Why I hated calculus but love statistics*

1 What's the Point?

2 Descriptive Statistics: *Who was the best baseball player of all time?*
Appendix to Chapter 2

3 Deceptive Description: *"He's got a great personality!" and other true but grossly misleading statements*

4 Correlation: *How does Netflix know what movies I like?*
Appendix to Chapter 4

5 Basic Probability: *Don't buy the extended warranty on your $99 printer*

5½ The Monty Hall Problem

6 Problems with Probability: *How overconfident math geeks nearly destroyed the global financial system*

7 The Importance of Data: *"Garbage in, garbage out"*

8 The Central Limit Theorem: *The Lebron James of statistics*

9 Inference: *Why my statistics professor thought I might have cheated*
Appendix to Chapter 9

10 Polling: *How we know that 64 percent of Americans support the death penalty (with a sampling error ± 3 percent)*
Appendix to Chapter 10

11 Regression Analysis: *The miracle elixir*
Appendix to Chapter 11

12 Common Regression Mistakes: *The mandatory warning label*

13 Program Evaluation: *Will going to Harvard change your life?*

Conclusion: *Five questions that statistics can help answer*
Appendix: *Statistical software*
Notes
Acknowledgments
Index

Copyright
Also by Charles Wheelan

Introduction

Why I hated calculus but love statistics

I have always had an uncomfortable relationship with math. I don't like numbers for the sake of numbers. I am not impressed by fancy formulas that have no real-world application. I particularly disliked high school calculus for the simple reason that no one ever bothered to tell me why I needed to learn it. What is the area beneath a parabola? Who cares?

In fact, one of the great moments of my life occurred during my senior year of high school, at the end of the first semester of Advanced Placement Calculus. I was working away on the final exam, admittedly less prepared for the exam than I ought to have been. (I had been accepted to my first-choice college a few weeks earlier, which had drained away what little motivation I had for the course.) As I stared at the final exam questions, they looked completely unfamiliar. I don't mean that I was having trouble answering the questions. I mean that I didn't even recognize what was being asked. I was no stranger to being unprepared for exams, but, to paraphrase Donald Rumsfeld, I usually knew what I didn't know. This exam looked even more Greek than usual. I flipped through the pages of the exam for a while and then more or less surrendered. I walked to the front of the classroom, where my calculus teacher, whom we'll call Carol Smith, was proctoring the exam. "Mrs. Smith," I said, "I don't recognize a lot of the stuff on the test."

Suffice it to say that Mrs. Smith did not like me a whole lot more than I liked her. Yes, I can now admit that I sometimes used my limited powers as student association president to schedule all-school assemblies just so that Mrs. Smith's calculus class would be canceled. Yes, my friends and I did have flowers delivered to Mrs. Smith during class from "a secret admirer" just so that we could chortle away in the back of the room as she looked around in embarrassment. And yes, I did stop doing any homework at all once I got in to college.

So when I walked up to Mrs. Smith in the middle of the exam and said that the material did not look familiar, she was, well, unsympathetic. "Charles," she said loudly, ostensibly to me but facing the rows of desks to make certain that the whole class could hear, "if you had studied, the material would look a lot more familiar." This was a compelling point.

So I slunk back to my desk. After a few minutes, Brian Arbetter, a far better calculus student than I, walked to the front of the room and whispered a few things to Mrs. Smith. She whispered back and then a truly extraordinary thing happened. "Class, I need your attention," Mrs. Smith announced. "It appears that I have given you the second semester exam by mistake." We were far enough into the test period that the whole exam had to be aborted and rescheduled.

I cannot fully describe my euphoria. I would go on in life to marry a wonderful woman. We have three healthy children. I've published books and visited places like the Taj Mahal and Angkor Wat. Still, the day that my calculus teacher got her comeuppance is a top five life

moment. (The fact that I nearly failed the makeup final exam did not significantly diminish this wonderful life experience.)

The calculus exam incident tells you much of what you need to know about my relationship with mathematics—but not everything. Curiously, I loved physics in high school, even though physics relies very heavily on the very same calculus that I refused to do in Mrs. Smith's class. Why? *Because physics has a clear purpose.* I distinctly remember my high school physics teacher showing us during the World Series how we could use the basic formula for acceleration to estimate how far a home run had been hit. That's cool—and the same formula has many more socially significant applications.

Once I arrived in college, I thoroughly enjoyed probability, again because it offered insight into interesting real-life situations. In hindsight, I now recognize that it wasn't the math that bothered me in calculus class; it was that no one ever saw fit to explain the point of it. If you're not fascinated by the elegance of formulas alone—which I am most emphatically not—then it is just a lot of tedious and mechanistic formulas, at least the way it was taught to me.

That brings me to statistics (which, for the purposes of this book, includes probability). I love statistics. Statistics can be used to explain everything from DNA testing to the idiocy of playing the lottery. Statistics can help us identify the factors associated with diseases like cancer and heart disease; it can help us spot cheating on standardized tests. Statistics can even help you win on game shows. There was a famous program during my childhood called *Let's Make a Deal*, with its equally famous host, Monty Hall. At the end of each day's show, a successful player would stand with Monty facing three big doors: Door no. 1, Door no. 2, and Door no. 3. Monty Hall explained to the player that there was a highly desirable prize behind one of the doors—something like a new car—and a goat behind the other two. The idea was straightforward: the player chose one of the doors and would get the contents behind that door.

As each player stood facing the doors with Monty Hall, he or she had a 1 in 3 chance of choosing the door that would be opened to reveal the valuable prize. But *Let's Make a Deal* had a twist, which has delighted statisticians ever since (and perplexed everyone else). After the player chose a door, Monty Hall would open one of the two remaining doors, always revealing a goat. For the sake of example, assume that the player has chosen Door no. 1. Monty would then open Door no. 3; the live goat would be standing there on stage. Two doors would still be closed, nos. 1 and 2. If the valuable prize was behind no. 1, the contestant would win; if it was behind no. 2, he would lose. But then things got more interesting: Monty would turn to the player and ask whether he would like to change his mind and switch doors (from no. 1 to no. 2 in this case). Remember, both doors were still closed, and the only new information the contestant had received was that a goat showed up behind one of the doors that he didn't pick.

Should he switch?

The answer is yes. Why? That's in Chapter 5½.

The paradox of statistics is that they are everywhere—from batting averages to presidential polls—but the discipline itself has a reputation for being uninteresting and inaccessible. Many statistics books and classes are overly laden with math and jargon. Believe me, the technical

details are crucial (and interesting)—but it's just Greek if you don't understand the intuition. And you may not even care about the intuition if you're not convinced that there is any reason to learn it. Every chapter in this book promises to answer the basic question that I asked (to no effect) of my high school calculus teacher: *What is the point of this?*

This book is about the intuition. It is short on math, equations, and graphs; when they are used, I promise that they will have a clear and enlightening purpose. Meanwhile, the book is long on examples to convince you that there are great reasons to learn this stuff. *Statistics can be really interesting, and most of it isn't that difficult.*

The idea for this book was born not terribly long after my unfortunate experience in Mrs. Smith's AP Calculus class. I went to graduate school to study economics and public policy. Before the program even started, I was assigned (not surprisingly) to "math camp" along with the bulk of my classmates to prepare us for the quantitative rigors that were to follow. For three weeks, we learned math all day in a windowless, basement classroom (really).

On one of those days, I had something very close to a career epiphany. Our instructor was trying to teach us the circumstances under which the sum of an infinite series converges to a finite number. Stay with me here for a minute because this concept will become clear. (Right now you're probably feeling the way I did in that windowless classroom.) An infinite series is a pattern of numbers that goes on forever, such as 1 + ½ + ¼ + ⅛ . . . The three dots means that the pattern continues to infinity.

This is the part we were having trouble wrapping our heads around. Our instructor was trying to convince us, using some proof I've long since forgotten, that a series of numbers can go on forever and yet still add up (roughly) to a finite number. One of my classmates, Will Warshauer, would have none of it, despite the impressive mathematical proof. (To be honest, I was a bit skeptical myself.) How can something that is infinite add up to something that is finite?

Then I got an inspiration, or more accurately, the intuition of what the instructor was trying to explain. I turned to Will and talked him through what I had just worked out in my head. Imagine that you have positioned yourself exactly 2 feet from a wall.

Now move half the distance to that wall (1 foot), so that you are left standing 1 foot away.

From 1 foot away, move half the distance to the wall once again (6 inches, or ½ a foot). And from 6 inches away, do it again (move 3 inches, or ¼ of a foot). Then do it again (move 1½ inches, or ⅛ of a foot). And so on.

You will gradually get pretty darn close to the wall. (For example, when you are 1/1024th of an inch from the wall, you will move half the distance, or another 1/2048th of an inch.) But you will never hit the wall, because by definition each move takes you only half the remaining distance. In other words, you will get infinitely close to the wall but never hit it. If we measure your moves in feet, the series can be described as 1 + ½ + ¼ + ⅛ . . .

Therein lies the insight: Even though you will continue moving forever—with each move taking you half the remaining distance to the wall—the total distance you travel can never be more than 2 feet, which is your starting distance from the wall. For mathematical purposes, the total distance you travel can be approximated as 2 feet, which turns out to be very handy for

computation purposes. A mathematician would say that the sum of this infinite series 1 ft + ½ ft + ¼ ft + ⅛ ft . . . converges to 2 feet, which is what our instructor was trying to teach us that day.

The point is that I convinced Will. I convinced myself. I can't remember the math proving that the sum of an infinite series can converge to a finite number, but I can always look that up online. And when I do, it will probably make sense. In my experience, the intuition makes the math and other technical details more understandable—but not necessarily the other way around.

The point of this book is to make the most important statistical concepts more intuitive and more accessible, not just for those of us forced to study them in windowless classrooms but for anyone interested in the extraordinary power of numbers and data.

Now, having just made the case that the core tools of statistics are less intuitive and accessible than they ought to be, I'm going to make a seemingly contradictory point: Statistics can be *overly accessible* in the sense that anyone with data and a computer can do sophisticated statistical procedures with a few keystrokes. The problem is that if the data are poor, or if the statistical techniques are used improperly, the conclusions can be wildly misleading and even potentially dangerous. Consider the following hypothetical Internet news flash: *People Who Take Short Breaks at Work Are Far More Likely to Die of Cancer.* Imagine that headline popping up while you are surfing the Web. According to a seemingly impressive study of 36,000 office workers (a huge data set!), those workers who reported leaving their offices to take regular ten-minute breaks during the workday were 41 percent more likely to develop cancer over the next five years than workers who don't leave their offices during the workday. Clearly we need to act on this kind of finding—perhaps some kind of national awareness campaign to prevent short breaks on the job.

Or maybe we just need to think more clearly about what many workers are doing during that ten-minute break. My professional experience suggests that many of those workers who report leaving their offices for short breaks are huddled outside the entrance of the building smoking cigarettes (creating a haze of smoke through which the rest of us have to walk in order to get in or out). I would further infer that it's probably the cigarettes, and not the short breaks from work, that are causing the cancer. I've made up this example just so that it would be particularly absurd, but I can assure you that many real-life statistical abominations are nearly this absurd once they are deconstructed.

Statistics is like a high-caliber weapon: helpful when used correctly and potentially disastrous in the wrong hands. This book *will not* make you a statistical expert; it *will* teach you enough care and respect for the field that you don't do the statistical equivalent of blowing someone's head off.

This is not a textbook, which is liberating in terms of the topics that have to be covered and the ways in which they can be explained. *The book has been designed to introduce the statistical concepts with the most relevance to everyday life*. How do scientists conclude that something causes cancer? How does polling work (and what can go wrong)? Who "lies with

statistics," and how do they do it? How does your credit card company use data on what you are buying to predict if you are likely to miss a payment? (Seriously, they can do that.)

If you want to understand the numbers behind the news and to appreciate the extraordinary (and growing) power of data, this is the stuff you need to know. In the end, I hope to persuade you of the observation first made by Swedish mathematician and writer Andrejs Dunkels: It's easy to lie with statistics, but it's hard to tell the truth without them.

But I have even bolder aspirations than that. I think you might actually enjoy statistics. The underlying ideas are fabulously interesting and relevant. The key is to separate the important ideas from the arcane technical details that can get in the way. That is Naked Statistics.

CHAPTER 1

What's the Point?

I've noticed a curious phenomenon. Students will complain that statistics is confusing and irrelevant. Then the same students will leave the classroom and happily talk over lunch about batting averages (during the summer) or the windchill factor (during the winter) or grade point averages (always). They will recognize that the National Football League's "passer rating"—a statistic that condenses a quarterback's performance into a single number—is a somewhat flawed and arbitrary measure of a quarterback's game day performance. The same data (completion rate, average yards per pass attempt, percentage of touchdown passes per pass attempt, and interception rate) could be combined in a different way, such as giving greater or lesser weight to any of those inputs, to generate a different but equally credible measure of performance. Yet anyone who has watched football recognizes that it's handy to have a single number that can be used to encapsulate a quarterback's performance.

Is the quarterback rating perfect? No. Statistics rarely offers a single "right" way of doing anything. Does it provide meaningful information in an easily accessible way? Absolutely. It's a nice tool for making a quick comparison between the performances of two quarterbacks on a given day. I am a Chicago Bears fan. During the 2011 playoffs, the Bears played the Packers; the Packers won. There are a lot of ways I could describe that game, including pages and pages of analysis and raw data. But here is a more succinct analysis. Chicago Bears quarterback Jay Cutler had a passer rating of 31.8. In contrast, Green Bay quarterback Aaron Rodgers had a passer rating of 55.4. Similarly, we can compare Jay Cutler's performance to that in a game earlier in the season against Green Bay, when he had a passer rating of 85.6. That tells you a lot of what you need to know in order to understand why the Bears beat the Packers earlier in the season but lost to them in the playoffs.

That is a very helpful synopsis of what happened on the field. Does it simplify things? Yes, that is both the strength and the weakness of any descriptive statistic. One number tells you that Jay Cutler was outgunned by Aaron Rodgers in the Bears' playoff loss. On the other hand, that number won't tell you whether a quarterback had a bad break, such as throwing a perfect pass that was bobbled by the receiver and then intercepted, or whether he "stepped up" on certain key plays (since every completion is weighted the same, whether it is a crucial third down or a meaningless play at the end of the game), or whether the defense was terrible. And so on.

The curious thing is that the same people who are perfectly comfortable discussing statistics in the context of sports or the weather or grades will seize up with anxiety when a researcher starts to explain something like the Gini index, which is a standard tool in economics for measuring income inequality. I'll explain what the Gini index is in a moment, but for now *the most important thing to recognize is that the Gini index is just like the passer rating*. It's a handy tool for collapsing complex information into a single number. As such, it has the

strengths of most descriptive statistics, namely that it provides an easy way to compare the income distribution in two countries, or in a single country at different points in time.

The Gini index measures how evenly wealth (or income) is shared within a country on a scale from zero to one. The statistic can be calculated for wealth or for annual income, and it can be calculated at the individual level or at the household level. (All of these statistics will be highly correlated but not identical.) The Gini index, like the passer rating, has no intrinsic meaning; it's a tool for comparison. A country in which every household had identical wealth would have a Gini index of zero. By contrast, a country in which a single household held the country's entire wealth would have a Gini index of one. As you can probably surmise, the closer a country is to one, the more unequal its distribution of wealth. The United States has a Gini index of .45, according to the Central Intelligence Agency (a great collector of statistics, by the way).[1] So what?

Once that number is put into context, it can tell us a lot. For example, Sweden has a Gini index of .23. Canada's is .32. China's is .42. Brazil's is .54. South Africa's is .65. [*] As we look across those numbers, we get a sense of where the United States falls relative to the rest of the world when it comes to income inequality. We can also compare different points in time. The Gini index for the United States was .41 in 1997 and grew to .45 over the next decade. (The most recent CIA data are for 2007.) This tells us in an objective way that while the United States grew richer over that period of time, the distribution of wealth grew more unequal. Again, we can compare the changes in the Gini index across countries over roughly the same time period. Inequality in Canada was basically unchanged over the same stretch. Sweden has had significant economic growth over the past two decades, but the Gini index in Sweden actually fell from .25 in 1992 to .23 in 2005, meaning that Sweden grew richer *and* more equal over that period.

Is the Gini index the perfect measure of inequality? Absolutely not—just as the passer rating is not a perfect measure of quarterback performance. But it certainly gives us some valuable information on a socially significant phenomenon in a convenient format.

We have also slowly backed our way into answering the question posed in the chapter title: What is the point? The point is that statistics helps us process data, which is really just a fancy name for information. Sometimes the data are trivial in the grand scheme of things, as with sports statistics. Sometimes they offer insight into the nature of human existence, as with the Gini index.

But, as any good infomercial would point out, *That's not all!* Hal Varian, chief economist at Google, told the *New York Times* that being a statistician will be "the sexy job" over the next decade.[2] I'll be the first to concede that economists sometimes have a warped definition of "sexy." Still, consider the following disparate questions:

How can we catch schools that are cheating on their standardized tests?

How does Netflix know what kind of movies you like?

How can we figure out what substances or behaviors cause cancer, given that we cannot conduct cancer-causing experiments on humans?

Does praying for surgical patients improve their outcomes?