

# THE CURIOUS JOURNALIST'S GUIDE TO DATA

JONATHAN  
STRAY

# Table of Contents

Introduction	0
Dedication	1
Introduction	2
Quantification	3
The Quantities of Everyday Language	3.1
Counting Race	3.2
The Problem of What to Count	3.3
Sampling and Quantified Error	3.4
The Problem of Measurement Error	3.5
Quantification Is Representation	3.6
Analysis	4
Did the Policy Work?	4.1
Accounting for Chance	4.2
Counting Possible Worlds	4.3
Arguing From the Odds	4.4
Statistical Inference	4.5
What Would Have Happened Anyway?	4.6
Causal Models	4.7
Truth by Elimination	4.8
Communication	5
Perception	5.1
Representation	5.2
Examples Trump Statistics	5.3
Who Is in the Data?	5.4
Communicating Uncertainty	5.5
Prediction	5.6
Going Further	6
Footnotes	7
Citations	8

# The Curious Journalist's Guide to Data

## Dedication

For every journalist who has ever thought they're bad at math. What if you're wrong?

## Acknowledgments

Thank you to the Tow Center for Digital Journalism for the fellowship that supported the writing of this work. I could not have done this otherwise. I'm indebted to Mark Hansen for reading not one but two long drafts and providing expansive feedback. Andrew Gelman kindly reviewed the "Analysis" chapter and really shaped my thinking on causation. Kenneth Prewitt read the material on census and race with an expert eye; any remaining blunders are my own. I'm indebted to research directors Taylor Owen and Claire Wardle for their patient efforts as they shepherded me through the process over nearly two years. I'm deeply grateful to Emily Bell for her support over the years, and the fantastic opportunity to teach at Columbia. My warmest shout-out to the students of my Frontiers of Computational Journalism course, who taught me what it is to teach—and sometimes schooled me with their own work. You've been more influential than you know. And thank you to Sara for helping me find the book's title.

*March 2016*



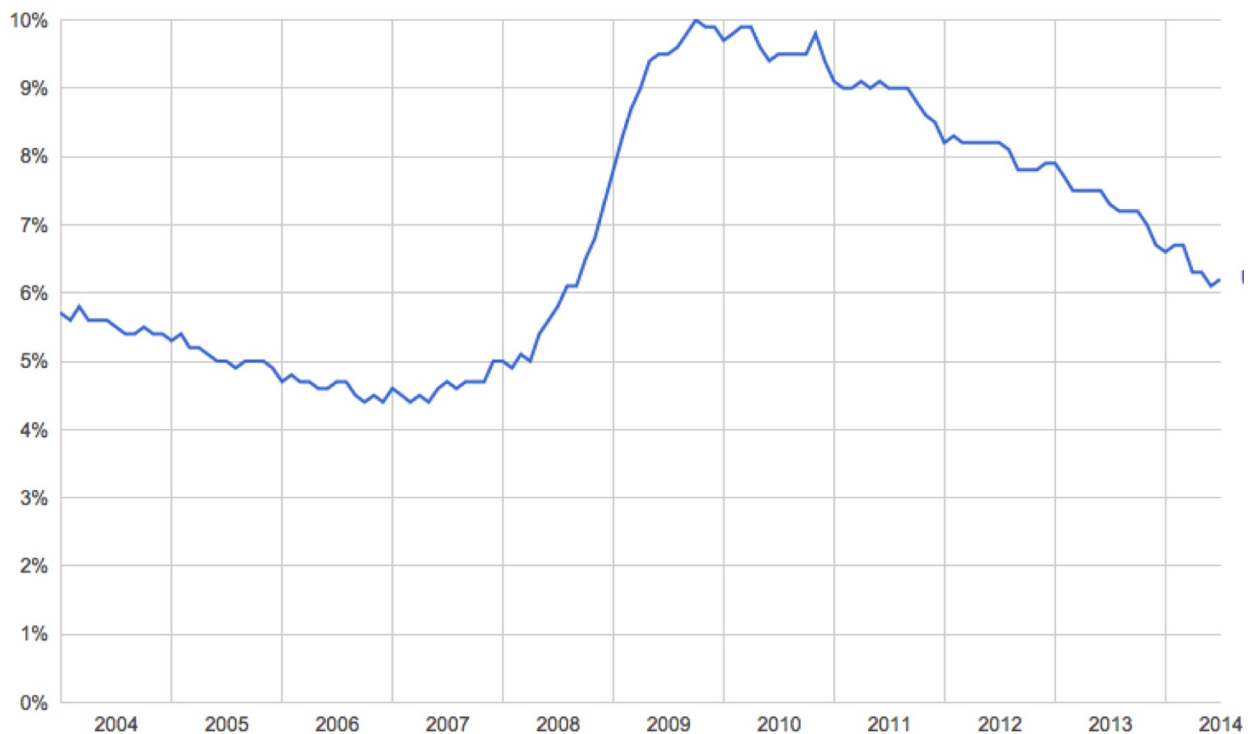
# Introduction

This is a book about using data in journalism, but it's not a particularly practical book. Instead it's for the curious, for those who wonder about the deep ideas that hold everything together. Some of these ideas are very old, some have emerged in just the last few decades, and many of them have come together to create the particularly twenty-first-century practice of data journalism.

We'll cover some of the mathy parts of statistics, but also the difficulty of taking a census of race and the cognitive psychology of probabilities. We'll trace where data comes from, what journalists do with it, and where it goes after—and try to understand the possibilities and limitations. Data journalism is as interdisciplinary as it gets, which can make it difficult to assemble all the pieces you need. This is one attempt.

There are few equations and no code in this book, and I don't assume you know anything about math. But I am assuming you want to know, so I'm going to develop some key ideas from the ground up. Or maybe you've studied a technical field and you are just coming into journalism, in which case I hope this book helps you understand how your skills apply. This is a framework, a collection of big ideas journalists can steal from other fields. I want to give a foothold into statistical analysis in all its nerdy splendor, but equally show how ethnography can help you interpret crime figures.

We're going to look at data a lot more closely than you might be used to. Consider this graph of the U.S. unemployment rate over the last 10 years. There is a whole world just beneath the surface of this image.



*From the U.S. Bureau of Labor Statistics.*

It's clear that a lot of people lost their jobs after the 2008 financial crash. You can read this chart and say how many: The unemployment rate went up by 5 percent. This is a very ordinary, very reasonable way of talking about this data, exactly the sort of thing that should pop into your head when you see this image. We're going to look deeper.

Where did these numbers come from? What do they actually count? What can the journalist say about this data, in light of recent history? What should the audience do after seeing it? Why do we believe charts like this, and should we? How is an unemployment chart any better, or different, than just asking people about their post-crash lives?

What's the data really doing for us here?

This book is about bringing the quantitative tradition into journalism. Data is not just numbers, but numbers were the first form of data. The very first writing systems were used for accounting, long before they were sophisticated enough for language.<sup>1</sup> At that time the rules of addition must have seemed incredibly arcane (in base 60, at first!), and it must have been a powerful trick to be able to tell in advance how many stones you would need for a building of a certain size. There is no doubt that numbers, like words, are a type of practical magic, and counting is the foundation of data work to this day. But you already know how to count. So we're mostly going to talk about ideas that were developed during The Enlightenment, then massively refined and expanded in the twentieth century with modern statistics and computers.

We'll need to go well outside of statistics to make any sense of things. I've been raiding psychology and social science and ethnography, and further places too like intelligence analysis and the neurobiology of vision. I've been collecting pieces, hoping to use data more thoughtfully and effectively in my journalism work. I've tried to organize the things that can be said into three parts: Quantification is what makes data, then the journalist analyzes it, then the result is communicated to the audience. This process creates "stories," the central products of journalism.

In journalism, a story is a narrative that is not only true but interesting and relevant to the intended audience. Data journalism is different from pure statistical analysis—if there is such a thing—because we need culture, law, and politics to tell us what data matters and how. A procurement database may tell us that the city councilor has been handing out lucrative contracts to his brother. But this is interesting only if we understand this sort of thing as "corruption" and we've decided to look for it. A sports journalist might look for entirely different stories in the same data, such as whether or not the city is actually going to build that proposed new stadium. The data alone doesn't determine the story. But the story still has to be true, and hopefully also thorough and fair. What exactly that means isn't always obvious. The relationship between story, data, culture, and truth is one of the key problems of twenty-first-century journalism.<sup>1</sup>

The process of quantification, analysis, and communication is a cycle. After communicating a result you may realize that you want a different analysis of the same data, or different data entirely. You might end up repeating this process many times before anything is ever published, exploring the data and communicating primarily to yourself and your colleagues to find and shape the story. Or these steps might happen for each of many stories in a long series, with feedback from the audience directing the course of your reporting. And somewhere, at some point, the audience acts on what you have communicated. Otherwise, journalism would have no effect at all.

Data begins with quantification. Data is not something that exists in nature, and unemployed people are a very different thing than unemployment data. What is counted and how?

There are at least six different ways that the U.S. government counts who is unemployed, which give rise to data sets labeled U1 to U6.<sup>2</sup> The official unemployment rate—the government calls one of them "official"—is known as U3. But U3 does not count people who gave up looking for a job, as U4 does, or people who hold part-time jobs because they can't get a full-time job, as U6 does.

And this says nothing about how these statistics are actually tabulated. No one goes around asking every single American about their employment status every single month. The official numbers are not "raw" counts but must be derived from other data in a vast and

sophisticated ongoing estimation process based on random sampling. Unemployment figures, being estimates, have statistical estimation error—far more error than generally realized. This makes most stories about short-term increases or decreases irrelevant.<sup>3</sup>

There is a complex relationship between the idea conveyed by the words “unemployment rate” and the process that produces a particular set of numbers. Normally all of this is backstage, hidden behind the chart. It’s the same for any other data. Data is created. It is a record, a document, an artifact, dripping with meaning and circumstance. A machine recorded a number at some point on some medium, or a particular human on a particular day made a judgment that some aspect of the world was this and not that, and marked a 0 or a 1. Even before that, someone had to decide that some sort of information was worth recording, had to conceive of the categories and meanings and ways of measurement, and had to set up the whole apparatus of data production.<sup>ii</sup>

Data production is an elaborate process involving humans, machines, ideas, and reality. It is social, physical, and specific to time and place. I’m going to call this whole process “quantification,” a word which I’ll use to include everything from dreaming up what should be counted to wiring up sensors.

If quantification turns the world into data, analysis tells us what the data means. Here is where journalism leans most heavily on traditional mathematical statistics. If you’ve found statistics difficult to learn, it’s not your fault. It has been terribly taught.<sup>4</sup> Yet the underlying ideas are beautiful and sensible. These foundational principles lead to certain rules that guide our search for truth, and we want those rules. It is hard to forgive arithmetic errors or a reporter’s confused causality. Journalism can demand deep and specific technical knowledge. It’s no place for people who want to avoid math.

Suppose you want to know if the unemployment rate is affected by, say, tax policy. You might compare the unemployment rates of countries with different tax rates. The logic here is sound, but a simple comparison is wrong. A great many things can and do affect the unemployment rate, so it’s difficult to isolate just the effect of taxes. Even so, you can build statistical models to help you guess what the unemployment rate *would have been* if all factors other than tax policy were the same between countries. We’re now talking about imaginary worlds, derived from the real through force of logic. That’s a tricky thing—not always possible, and not always defensible even when formally possible. But we do have hundreds of years of guidance to help us.

Journalists are not economists, of course. They’re not really specialists of any kind, especially if journalism is all they have studied and practiced. We already have economists, epidemiologists, criminologists, climatologists, and on and on. But journalists need to understand the methods of any field they touch, or they will be unable to tell good work from bad. They won’t know which analyses are worth repeating. Even worse, they will not

understand which data matters. And, increasingly, journalists are attempting their own analyses when they discover that the knowledge they want does not yet exist. Journalists aren't scientists, but they need to understand what science knows about evidence and inference.

There are few outright equations in this book, but it is a technical book. I use standard statistical language and try to describe concepts faithfully but mostly skip the formal details. Whenever you see a word in *italics* that means you can go look it up elsewhere. Each technical term is a gateway to whole worlds of specialized knowledge. I hope this book gives you a high-level view of how statistical theory is put together, so you'll know what you're trying to do and where you might look for the appropriate pieces.

After analysis comes communication. This makes journalism different from scholarship or science, or any field that produces knowledge but doesn't feel the compulsion to tell the public about it in an understandable way. Journalism is *for* the audience—which is often a very broad audience, potentially millions of people.

Communication depends on human culture and cognition. A story includes an unemployment chart because it's a better way of communicating changes in the unemployment rate than a table of numbers, which is true because human eyes and brains process visual information in a certain way. Your visual system is attuned to the orientation of lines, which allows you to perceive trends without conscious effort. This is a remarkable fact which makes data visualization possible! And it shows that data journalists need to understand quantitative cognition if they want to communicate effectively.

From experience and experiments we know quite a lot about how minds work with data. Raw numbers are difficult to interpret without comparisons, which leads to all sorts of normalization formulas. Variation tends to get collapsed into stereotypes, and uncertainty tends to be ignored as we look for patterns and simplifications. Risk is personal and subjective, but there are sensible ways to compare and communicate odds.

But more than these technical concerns is the question of what is being said about whom. Journalism is supposed to reflect society back to itself, but who is the "we" in the data? Certain people are excluded from any count, and astonishing variation is abstracted into uniformity. The unemployment rate reduces each voice to a single bit: are you looking for work, yes/no? A vast social media data set seems like it ought to tell us deep truths about society, but it cannot say anything about the people who don't post, or the things they don't post about. Omniscience sounds fantastic, but data is a map and not the territory.

And then there's the audience. What someone understands when they look at the data depends on what they already believe. If you aren't unemployed yourself, you have to rely on some image of "unemployed person" to bring meaning to the idea of an unemployment rate. That image may be positive or negative, it may be justified or untrue, but you have to fill

in the idea of unemployment with *something* to make any sense at all of unemployment statistics. Data can demolish or reinforce stereotypes, so it's important for the journalist to be aware that these stereotypes are in play. That is one reason why it's not enough for data to be presented "accurately." We have to ask what the recipient will end up believing about the world, and about the *people* represented by the data. Often, data is best communicated by connecting it to stories from the individual lives it represents.

We're not quite done. I want action. Someone eventually has to act on what they've learned if journalism is going to mean anything at all, and action is a powerfully clarifying perspective. Knowing the unemployment rate is interesting. Much better is knowing that a specific plan would plausibly create jobs. This sort of deep research will usually be done by specialists, but journalists have to understand enough to act as a communicator and an independent check. As a media professional, a journalist has both the power and responsibility to decide what is worth repeating.

Data cannot tell us what to do, but it can sometimes tell us about consequences. The twentieth century saw great advances in our understanding of causality and prediction. But prediction is very hard. Most things can't be predicted well, for fundamental reasons such as lack of data, intrinsic randomness, free will, or the butterfly effect. These are profound limits to what we can know about the future. Yet where prediction is possible, there is convincing evidence that data is essential. Purely qualitative methods, no matter how sophisticated, just don't seem to be as accurate. Statistical methods are essential for journalism that asks what will happen, what should be done, or how best to do it.

This doesn't mean we can just run the equations forward and read off what to do. We've seen that dream before. At an individual level, the ancient desire for universal quantification can be a source of mathematical inspiration. Leibniz dreamed of an unambiguous language of "universal character." Three centuries later, the failure of the symbolic logic paradigm in artificial intelligence finally showed how impractical that is, but the exercise was enormously productive. The desire for universal quantification hasn't worked out quite so well at a societal level. Every authoritarian planner dreams of utopia, but totalitarian technocratic visions have been uniformly disastrous for the people living in them. A fully quantified social order is an insult to freedom, and there are good reasons to suspect such systems will always be defeated by their rigidity.<sup>5</sup> Questions of action can hone and refine data work, but actual action—making a choice and doing—requires practical knowledge, wisdom, and creativity. The use of statistics in journalism, like the use of statistics in general, will always involve artistry.

All of this is implicit in every use of data in journalism. All of it is just below the surface of an unemployment chart in the news, to say nothing of the dazzling visualizations that journalists now create. Journalism depends on what we have decided to count, the techniques used to

interpret those counts, how we decide to show the results, and what happens after we do. And then the world changes, and we report again.

# Quantification

*The mathematical modeling tools we employ at once extend and limit our ability to conceive the world.* - David Hestenes<sup>6</sup>

There were no Hispanics living in the United States before 1970. At least, there weren't according to the census. There couldn't be, because the census form did not include "Hispanic" or "Latino" or anything like it.<sup>iii</sup>

Actually there were about nine million Hispanics living in the country by 1970.<sup>7</sup> In many ways the lack of census data made them invisible. You couldn't say with certainty where they were living. It would have been difficult to know how the health, education, and income of Hispanic families compared to other families, much less contemplate ways to close the gaps. You wouldn't even know how many people might be affected if you did.

Quantification is the process that creates data. You can only measure what you can conceive. That's the first challenge of quantification. The next challenge is actually measuring it, and knowing that you measured it accurately. Data is only useful because it represents the world, but that link can be fragile. At some point, some person or machine counted or measured or categorized, and recorded the result. The whole process has to work just right, and our understanding of exactly how it all works has to be correct, or the data won't be meaningful.

Sometimes this is not a simple thing to do. It seems clear enough how to quantify the number of cars sold or the amount of grain exported, where counting has the feel of something objective and definite. But journalists are interested in many other things where the proper relationship between the words, the numbers, and the world is much less clear.

Are mass shootings more or less common today than 10 years ago? What fraction of the population is Hispanic? How many people suffer from depression? These seem like questions that counting can answer, but "mass shootings," "Hispanics," and "depression" are not easy things to count. Who, precisely, counts as depressed? And how would you determine the number of depressed people in the entire country?

Quantification is a problem without a home. Statisticians and computer scientists do not normally spend a lot of time asking how data came to be. Actually, their methods are powerful precisely because they are abstract. Physicists and engineers were the first to think seriously about quantification, and they have carefully developed the processes of measurement over many centuries. Even in such "hard" disciplines there are many choices that must be made about what gets measured, but these fields usually only deal with quantities that can be expressed in the units of physics. Econometrics broadened the



horizons, but it is psychologists and social scientists who have thought most deeply about the quantification of people and societies, the sorts of quantifications that are often most interesting and most vexing to a journalist.<sup>iv</sup>

I'm going to try to give the flavor of the problems of quantification with two examples: recording someone's race in a database and estimating the monthly unemployment rate. The first is a parable about the difficulty of categories. The second is a tour through the beautiful ideas of random sampling and quantified uncertainty so central to modern statistical work. But before we can get there, we have to talk about what makes something "quantitative" at all.

## The Quantities of Everyday Language

Quantity is an ancient idea, so ancient that it appears at the core of every human language. Words like “less” and “every” are obviously quantitative, and lead to more complex concepts like “trend” and “significant.” Quantitative thinking starts with recognizing when you are talking about quantities.

Spot the quantitative ideas in this sentence from the article “Anti-Intellectualism is Killing America,” which appeared in *Psychology Today*:

In a country where a sitting congressman told a crowd that evolution and the Big Bang are “lies straight from the pit of hell,” where the chairman of a Senate environmental panel brought a snowball into the chamber as evidence that climate change is a hoax, where almost one in three citizens can’t name the vice president, it is beyond dispute that critical thinking has been abandoned as a cultural value. > 8

This is pure cultural critique, and we could take it many different ways. We could read this sentence as a rant, a plea, an affirmation, a provocation, a list of examples, or any other type of expression. Maybe it’s art. But journalism is traditionally understood as “nonfiction,” so let’s take this at face value and ask whether it’s true.

I see an empirical and quantitative claim at the heart of the phrase “critical thinking has been abandoned as a cultural value.” it’s empirical because it speaks about something that is happening in the world, something with observable consequences. it’s quantitative because the word “abandoned” speaks about comparing the amount of something at two different times. Something we never had can’t be abandoned.

For at least two points in time we need to decide whether or not “critical thinking is a cultural value.” This is the moment of quantification. “Abandoned” might have an all-or-nothing flavor, but it’s probably a lot more reasonable to define shades of gray based on the number of people and institutions that are embodying the value of critical thinking; or perhaps it makes sense to look at how many acts of critical thinking are occurring. Of course “critical thinking” is not an easy thing to pin down but if we choose any definition at all we are literally deciding which things “count” as critical thinking. The next step is to come up with a practical plan to count those things. If we can’t or won’t count in practice, there’s no quantitative way to test this claim against reality. it’s not that the sentence would then mean nothing, it’s just that its meaning couldn’t be evaluated by comparing the words with the world in a yes/no kind of way.

One way or another, testing the claim that “critical thinking has been abandoned as a cultural value” demands that we count something at two different points in time and look for a drop in the number. There are surely fights waiting to happen over what should be counted, whether it was correctly counted, and the numerical threshold for “abandoned.” But if you’re willing to make some choices, you can go out and find relevant facts. This is what the author’s given us:

- a sitting congressman told a crowd that evolution and the Big Bang are “lies straight from the pit of hell”
- the chairman of a Senate environmental panel brought a snowball into the chamber as evidence that climate change is a hoax
- almost one in three citizens can’t name the vice president

Even if these were all good examples of a failure of “critical thinking,” they still wouldn’t be good evidence for the idea that critical thinking has been abandoned. The problem is that the author is trying to say something about a very large group of people. These examples would need to be representative. Are these failures of critical thinking typical of the whole society? It seems just as easy to come up with counterexamples. Yeah, someone brought a snowball into Congress to argue against climate change, but the EPA also recently decided to start regulating carbon dioxide as a pollutant. That’s evidence against the representativeness of the author’s examples, but of course you could dig up a million more examples on each side. That’s where counting gets interesting: it’s a systematic way to grasp the whole of something, which can lead to much stronger statements.

That’s the logic behind historian G. Kitson Clark’s advice for making generalizations:

Do not guess; try to count. And if you cannot count, admit that you are guessing.<sup>9</sup>

The fact that “one in three citizens can’t name the vice president” is closer to the sort of evidence we need. This statement generalizes in a way that individual examples can’t, because it makes a claim about all U.S. citizens. It doesn’t matter how many people I can name who know who the vice president is, because we know (by counting) that there are 100 million who cannot. But this still only addresses one point in time. Were things better before? Was there any point in history where more than two-thirds of the population could name the vice-president? We don’t know.

In short, the evidence in this sentence is not the right type. The word “abandoned” has embedded quantitative concepts that are not being properly handled. We need something tested or measured or counted across the entire culture at two different points in time, and we don’t have that—none of which makes this a “bad” piece of writing. It might provoke the reader to think about the value of critical thinking. It might be emotionally resonant. It might draw attention to important examples. It might even be persuasive. Whether it’s good or not

depends on what you want it to do. But in terms of empirical claims and the evidence provided for them, this is a weak argument. It doesn't respect the quantitative structure of the language it uses.

Many words have quantitative aspects. Words like "all," "every," "none," and "some" are so explicitly quantitative that they're called quantifiers in mathematics. Comparisons like "more" and "fewer" are clearly about counting, but much richer words like "better" and "worse" also imply counting or measuring at least two things. There are words that compare different points in time, like "trend," "progress," and "abandoned." There are words that imply magnitudes such as "few," "gargantuan," and "scant." A series of Greek philosophers, long before Christ, showed that the meanings of "if," "then," "and," "or," and "not" could be captured symbolically as *propositional logic*. To be sure, all of these words have meanings and resonances far beyond the mathematical. But they lose their central meaning if the quantitative core is ignored.

We're really taking language apart here, and no one could make it through a day if they had to fact check every sentence they read. Also, there are other ways of relating to a story. But this is a way of seeing that every journalist should have in their toolbox—and pass on to readers when helpful. The relation between words and numbers is of fundamental importance to the pursuit of truth. It tells you when you should be counting something.

## Counting Race

In 2004, the government of Florida drew up a list of felons who were ineligible to vote. It did this by matching names between a criminal records database and a registered voter database. The courts ordered that the list be released publicly, and shortly thereafter the *Sarasota Herald-Tribune* discovered that there were almost no Hispanics on the list.<sup>10</sup>

This seemed impossible. Hispanics made up 17 percent of the population but only one-tenth of 1 percent of the list; there were only 61 Hispanic people on the list of 47,763 names. At the time, Florida's Hispanic voters were mostly Cubans who supported the Republican Party. If they weren't on the list, they would be allowed to vote. There were accusations of politically motivated fraud.

More digging revealed that this was not actually a political maneuver but a data problem. In the state's voter database, Hispanic is a "race." In the criminal history database, Hispanic is an "ethnicity." The same information was conceived in two different ways, so it was recorded in two different fields in two different systems. To prevent false matches based on name alone, the government had chosen to match on name, date of birth, and "race" but not "ethnicity." Thus, Hispanic felons could never match Hispanic voters.<sup>11</sup>

Which database schema is correct? Is Hispanic an ethnicity or a race? This sounds like a cultural, social, or even philosophical question, but in this context it's really a question about the process of counting. After all, these databases are concrete objects, created by humans. At some point there was a decision that each person was, or was not, Hispanic, and this value was recorded in either the "race" or "ethnicity" column.

How do you assign a racial category to each person, or even decide what those categories should be? This is a problem that the U.S. Census has solved, for better or worse, for over 200 years.

Article I, Section 2 of the 1787 Constitution established the census and divided people into three categories: "free persons"; "Indians not taxed"; and "other persons," which really meant "slaves." Although aligned with race, these were also political categories because the census was created to apportion representatives and taxes between the states. Indians counted for neither representation nor taxes, while slaves were only counted as three-fifths of a person. This was the compromise between the slave and non-slave states that created the country. It seems insane now, but that's the history, and a reminder that the census is not an "objective" count but a bureaucratic process that generates data for specific purposes. Asking *why* the data was collected does not answer how it was collected, but it's often a big hint.

Over the next century it became possible for a person to be counted in many more different ways. The category of “free colored person” appeared in 1820. No one was interracial, according to the data, until the 1850 census added the category of “mulatto.” The 1890 census expanded into ethnicity and finer shades of black when it asked “whether white, black, mulatto, quadroon, octoroon, Chinese, Japanese, or Indian.”

Of course you could see people of all these types on city streets by then—but not in the official statistics until these additions. Categories were being added to better describe a reality that could already be perceived by other means. Which doesn’t make the categories reality. There were huge numbers of people who didn’t fit into any of these categories, like the Irish, who suffered intense racism in nineteenth-century America.

But a list of races doesn’t tell us how a person’s race was actually determined. In practice, a census enumerator visited each home and checked a box. For decades, enumerators were told to count someone as black if there was any degree of black ancestry, echoing the “one drop rule” of the Jim Crow era. Here’s how race was supposed to be quantified for the 1940 census:

## Personal Description

452. **Column 9. Sex.**—Write “M” for male, and “F” for female.

453. **Column 10. Color or Race.**—Write “W” for white; “Neg” for Negro; “In” for Indian; “Chi” for Chinese; “Jp” for Japanese; “Fil” for Filipino; “Hin” for Hindu; and “Kor” for Korean. For a person of any other race, write the race in full.

454. **Mexicans.**—Mexicans are to be regarded as white unless definitely of Indian or other nonwhite race.

455. **Negroes.**—A person of mixed white and Negro blood should be returned as a Negro, no matter how small the percentage of Negro blood. Both black and mulatto persons are to be returned as Negroes, without distinction. A person of mixed Indian and Negro blood should be returned as a Negro, unless the Indian blood very definitely predominates and he is universally accepted in the community as an Indian.

456. **Indians.**—A person of mixed white and Indian blood should be returned as Indian, if enrolled on an Indian Agency or Reservation roll; or if not so enrolled, if the proportion of Indian blood is one-fourth or more, or if the person is regarded as an Indian in the community where he lives. (See par. 455 for mixed Indian and Negro.)

457. **Mixed Races.**—Any mixture of white and nonwhite should be reported according to the nonwhite parent. Mixtures of nonwhite races should be reported according to the race of the father, except that Negro-Indian should be reported as Negro.

*Instructions for quantifying race and sex on the 1940 census.*<sup>12</sup>

It's not clear how census-takers were supposed to determine someone's ancestry going back generations, or how they applied this rule in practice, or if they even read the instructions—meaning that we don't know quite how to interpret the racial categories of the early twentieth-century census. If the collection method is obscure, so is the data.

Then things changed. In the mid-twentieth century there was a huge shift in the way race was counted, but not because of social or philosophical ideals. Instead the motive was statistical accuracy.

Close analysis of the 1940 census data suggested that the results were low by 3.6 percent, meaning millions of people had not been counted. The census was supposed to be a simple count, but the massive undercount proved that counting was anything but simple. And some

people were more undercounted than others: 13 percent of non-“white” people were missing from the census results.

There was clearly a racial bias in the census-taking process. It was soon discovered that census enumerators were having difficulty identifying American Indians in urban areas where they were mixed in with majority white populations. This proved that looking at someone didn't always provide an accurate impression of their race. To address this, the 1960 census used a different approach: People were simply asked what race they were.

If self-identification seems the obvious way to determine race, that's because we now understand race as an entanglement of identity, culture, and biology, as much social as genetic. But that is a late twentieth-century understanding. The census officials of the 1950s do not seem to have understood race this way; they simply wanted a more accurate count and took for granted that a person knows their own race.

There is something about self-identification that feels like a step forward in codifying race, a better way of making it visible in the aggregate. it's a more dignified approach. But it has its own serious limitations. it's not the data you need if you want to study race-linked genetic diseases or how people treat strangers differently based on skin color. We can think of race in many different ways, but the available data has no obligation to match our conceptions. If you want to know what the data really measures, the only thing that matters is how it was collected. Hence, the census up to 1950 counts something different than the census from 1960 onward, even though both call it “race.” How is it different? That depends on the question you wish to ask of the data.

Meanwhile, Hispanics had begun to make up a significant fraction of the U.S. population, and “Hispanic” finally appeared on census forms in 1970. Before that the census said nothing about how many Hispanic people lived in the country, where they lived, their incomes, or any of the other variables now routinely collected.

Things changed again in 1977 with a new set of federal government guidelines on the collection of race data, the infamous “Directive 15” from the Office of Management and Budget. This recommended dividing race into four categories: “American Indian or Alaska Native,” “Asian or Pacific Islander,” “Black,” and “White.” It also said “it is preferable to collect data on race and ethnicity separately” and defined ethnicity as “Hispanic origin” or “not of Hispanic origin.” The logic here is that Hispanics can be any race, such as Afro-Cubans. Which is great, except that about a third of all Hispanic people consider “Hispanic” to be a race, or at least they check “other race” on their census forms and write in “Hispanic” or “Mexican” or “Latina.”<sup>13</sup>

This is how Florida's criminal history database came to code Hispanics differently than Florida's voter registration database. The database of felons coded race according to federal standards, so race could only be white, black, Asian, American Indian, or unknown. Hispanic



was coded as an ethnicity, in a different field. Meanwhile, the voter registration database coded Hispanic as a race. A simple comparison on the “race” field failed, because race is not a simple thing to quantify.

If the federal racial categorization system feels a bit arbitrary, that’s because it is. Even its creators knew not to take it too seriously, writing, “These classifications should not be interpreted as being scientific or anthropological in nature.”<sup>14</sup> Nonetheless, all of the federal government’s race data includes these four master categories to this day. But many agencies also collect more detailed information on racial sub-categories. The census has long included a growing list of Asian races, and you’ve been able to write in any race you want since 1910.

The last major change to the race questions on the census came in 2000. Now you’re allowed to check multiple races on the census form, in addition to several possible choices for Hispanic ethnicity. The 2010 form looked like this:

→ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

**8. Is Person 1 of Hispanic, Latino, or Spanish origin?**

- No, not of Hispanic, Latino, or Spanish origin
- Yes, Mexican, Mexican Am., Chicano
- Yes, Puerto Rican
- Yes, Cuban
- Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.* ↴

**9. What is Person 1's race? Mark  one or more boxes.**

- White
- Black, African Am., or Negro
- American Indian or Alaska Native — *Print name of enrolled or principal tribe.* ↴

- |   |  |  |
|---|--|--|
| <input type="checkbox"/> Asian Indian   | <input type="checkbox"/> Japanese  | <input type="checkbox"/> Native Hawaiian       |
| <input type="checkbox"/> Chinese  | <input type="checkbox"/> Korean  | <input type="checkbox"/> Guamanian or Chamorro |
| <input type="checkbox"/> Filipino   | <input type="checkbox"/> Vietnamese  | <input type="checkbox"/> Samoan                |
| <input type="checkbox"/> Other Asian — <i>Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.</i> ↴ | <input type="checkbox"/> Other Pacific Islander — <i>Print race, for example, Fijian, Tongan, and so on.</i> ↴ |  |

- Some other race — *Print race.* ↴

On the 2010 census, 2.9 percent of the population identified as two or more races. This is nine million people who are expressing a type of racial identity which was invisible before we decided to count it.

## The Problem of What to Count

Quantification always involves complex choices, even in the hard sciences. Although friction is a basic force of classical physics, it comes from micro-interactions between surfaces that aren't fully understood. A high school physics textbook will tell you that we usually describe it with two numbers: the coefficient of static friction which is how hard you have to push to start sliding, and the coefficient of kinetic friction which is how hard you have to push to keep sliding. But more sophisticated measurements show that friction is actually quite a complex force. It also depends on velocity, and even on how fast you were sliding previously.<sup>15</sup> Anyone working with friction has to choose how to quantify it.

Race is even more difficult to quantify, as are a great many things of social interest. It's terribly easy to forget this complexity when you are looking at neat rows and columns of data.

A few years ago I worked on a story about gun violence. At the time there was a lot of popular discussion about "mass shooting" incidents, and whether they were or weren't on the rise. But what's a "mass shooting"? It seems like a single murder doesn't count, so how many people must be killed at once before it's "mass"? You have to answer this question before you can answer the question of whether such incidents are more or less common than before. I eventually chose four people as the minimum threshold for a mass shooting, because that's what the data I had used. The creators of that data chose four because this is how the FBI counts "mass murders," even though those aren't quite the same thing as "mass shootings." Responding to the interest in these events, the FBI later released its own data set of "active shooter" incidents, which it defined as "individuals actively engaged in killing or attempting to kill people in populated areas (excluding shootings related to gang or drug violence)."<sup>v</sup>

This is all somewhat arbitrary, and there is no "right" answer here. What you should count depends on what you care about, that is, it depends on the story you are attempting to tell. And after looking at the data you may realize that you want to count something else. Your initial story may turn out to be uninteresting, unfair, or just plain wrong.

It gets even trickier. Imagine tracking the prevalence of mental health issues such as "depression" or "borderline personality disorder," which are short names for evolving ideas about diseases. The complex diagnostic criteria for these conditions, which used to be printed in thick handbooks, define a quantification process. Or think of the police officer who must record if a particular incident is "sexual harassment" or not. It's easy to imagine that not

every officer will have the same idea of what sexual harassment means. This can make the data maddeningly hard to interpret, not to mention unfair. Small differences in counting technique can and do become the focus of intense arguments.

Still we find some way to count. A quantification process formalizes the act of counting or measuring or categorizing and attempts to apply it consistently across many situations. That's the whole point of standard units like meters and kilograms. But alas, many vital things do not have standard measures. How do we quantify more abstract concepts such as "educational attainment" or "quality of life" or "intelligence"?

In practice we end up replacing such rich concepts with much simpler proxies. We get "test scores" instead of "educational attainment" and "income" as a proxy for "quality of life," while "intelligence" is today measured by a battery of tests which assess many different cognitive skills. In experimental science this is called operationalizing a variable, a fancy name for picking a definition that's both analytically useful and practical enough to create data.

If you want to ask a question that only quantitative methods can answer, you have little choice but to make this switch from rich conception to repeatable measurement. But quantification can also force you to be clear. Trying to quantify might lead you to discover that you've been using certain words for a long time without really understanding what they mean—do you really know what "intelligence" is? Eventually a quantification of a thing can become the definition, as the IQ test did. This might be a clarifying improvement, or a narrowing of perception, or both. In any case, it is a choice that should be made consciously.

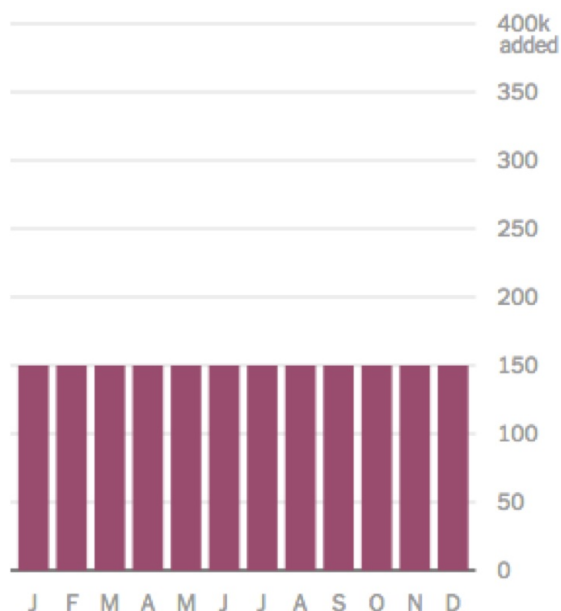
Usually there is some end goal, some purpose to collecting data, and you can ask whether any particular quantification method serves that purpose. And you can ask about the end purpose, too, the frame of the entire thing. Different quantification methods serve different stories.

## Sampling and Quantified Error

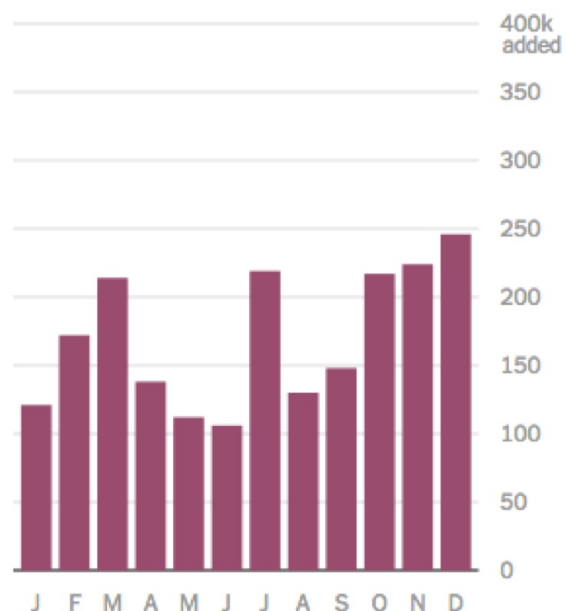
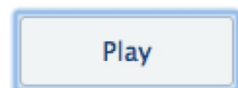
You should be skeptical of any headline that says the number of jobs in the United States has changed by fewer than about 105,000 since last month. That's because the monthly jobs growth estimate has a *margin of error* of about plus or minus 105,000.<sup>16</sup>

*The New York Times* made this point with an interactive graphic, showing how the uncertainty in employment figures can badly mislead us.

If job growth **were actually steady** over the last 12 months...



...the jobs report **could look like this:**



Did you infer a pattern from the chart on the right? If you did, you were reading too much into the jobs numbers. That chart should be a straight line, but we've added the same amount of sampling error that the jobs report has.

From *The New York Times*, 2014.<sup>17</sup>

Here, job growth was consistent at 150,000 new jobs each month, but the released figures show an upward trend just by chance. The unemployment rate calculated by the Bureau of Labor Statistics includes a fair amount of error due to random sampling, up to 105,000 jobs above or below the real value. Pressing "play" animates the right hand chart through endless possible scenarios with the same range of error. If you wait for a minute you can see cases

where job growth appears to have any trend you like. Because of these random errors, monthly changes typically mean less than we think they do. Long-term trends are much more reliable.

Political polls also have built-in error. If one candidate is ahead of the other 47 percent to 45 percent, but the margin of error is 5 percent, there is a pretty good chance that another identical poll will show the candidates the other way around. Pretty much any sort of public survey will have intrinsic error, and a reputable source will report the margin of error along with the results. The error of a measurement is a necessary part of understanding what that measurement means.

Maybe you've seen formulas for calculating the margin of error for a random sample, but rather than repeat those equations I want to give a sense of why we use random sampling at all and how it leads to quantified error. Expressing *how much* error there is may seem obvious now, but it was a key innovation in the history of statistics. There is a random sample in the Old Testament: "The people cast lots to bring one out of every ten of them to live in Jerusalem."<sup>vi</sup> It couldn't have been long before someone thought of counting by letting each of the chosen stand for 10, but millennia passed before anyone was able to estimate the accuracy of this process.

Sampling is basically a labor-saving device. The unemployment figures need to come out every month, but nobody is going to knock on your door 12 times a year to ask if you have a job. Instead the unemployment rate is calculated from the answers to two surveys: the Current Establishment Survey which samples businesses, and the Current Population Survey which samples households.<sup>vii</sup> 150,000 randomly chosen people each month,<sup>viii</sup> each of whom is eventually assigned to one of three categories: "employed," "unemployed," or "not in the labor force."<sup>18</sup> The fraction of "unemployed" people among those asked then stands in for the fraction of unemployed people in the whole country.

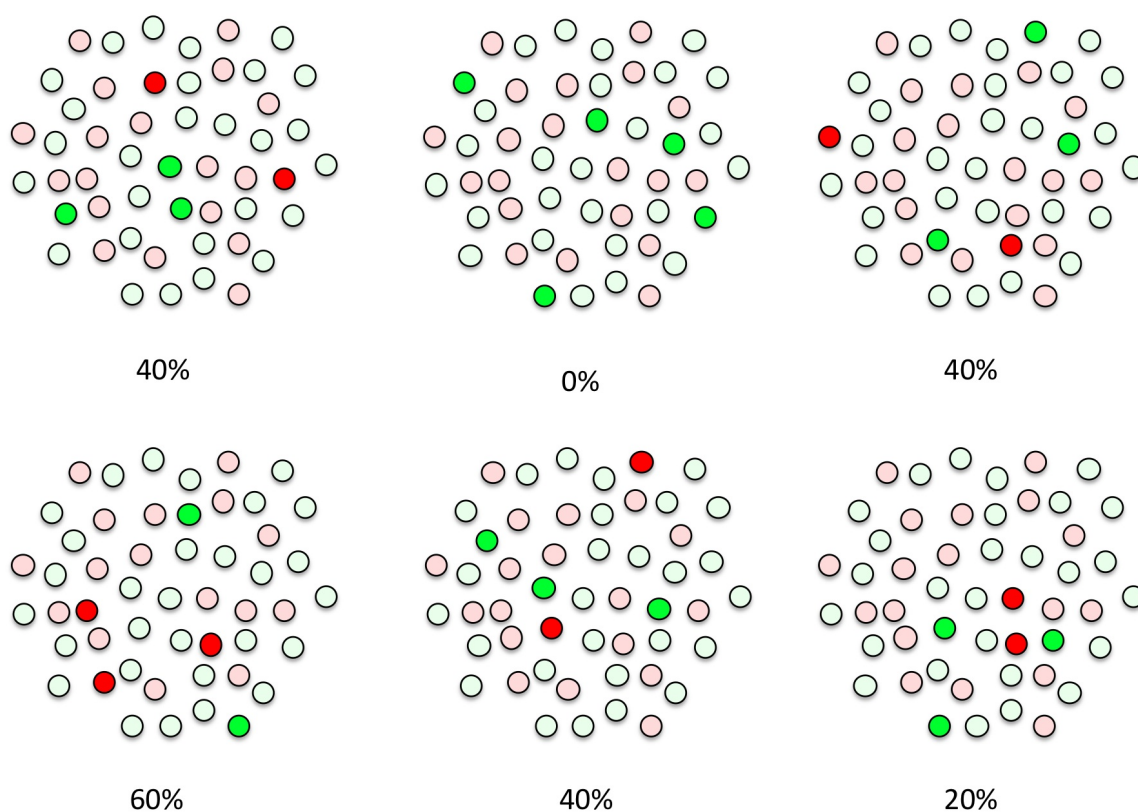
If this doesn't strike you as audacious, you've probably never thought about just what a poll claims to be able to do. Extrapolating from 150,000 people to 300,000,000 people means collecting information from one person in 2,000 then saying it speaks for the other 1,999. It's like asking only one person in each neighborhood whether he or she is employed.

Randomness is the key to this, because it makes over-representation by any one group extremely unlikely. It's possible that all the people who answer a random telephone poll might be unemployed just by chance, giving us a bad estimate. But that will happen rarely—essentially never in practice—and how else should we pick people? We could count through consecutive phone numbers instead, but that might only get us answers from a certain area. Or we could just go through our own contact lists, but that seems even less representative. Randomness is not subject to *selection bias* precisely because it has no relation to anything else. Even better, although any given sample will give us an estimate that is off by some



amount, the most common value is going to be the true value. Also, it's randomness that allows us to reason about what the error is. Instead of reasoning about the error of a single survey, which is unknowable, we can reason about the error of the sampling process across many different surveys. This is akin to saying that we can't know what the next roll of the die will be, but there is a one-sixth chance it will be a five.

Let's make the problem a little simpler and imagine that there are only 50 people in the whole country, and you've computed the unemployment rate by sampling five of them. You could have ended up with many different sets of five people in your sample had chance taken a different course, but there are a finite number of possibilities. Here are some of them, and the different unemployment rate estimates that each one would give you:



You can imagine drawing a picture of every possible set of names out of 50. You'll end up

with "50 choose 5" different sampling patterns, a number which is usually written  $\binom{50}{5}$ . You can get an actual number for this using the "choose" or "combinations" function of a scientific calculator or programming language, and it's 2,118,760, over two million. There are an awful lot of ways to pick five random things out of 50 possible things, and a hugely larger number of ways to pick 150,000 people out of 300,000,000, but we can count with simple formulas either way.

We can group all of these sampling patterns into six piles, according to how many people in each sample turned up unemployed, zero to five. This groups our answers into unemployment rates of 0/5, 1/5, 2/5, 3/5, 4/5, and 5/5, which is the same as 0%, 20%, 40%, 60%, 80%, and 100% unemployment. Because each possible sample—each set of five names—is equally likely, the size of each pile tells you your chances of getting a final estimate with that number of unemployed people. This is the key insight that will allow us to quantify how often we expect our unemployment estimate to be wrong, and by how much.

You don't actually need stacks of drawings to calculate the error of an unemployment estimate, because we can directly calculate the number of samples of each kind. For example, we can work out how many samples include exactly one unemployed person. Here there are 50 people, 20 of whom are unemployed. The total number of ways to choose five people from 50 so that exactly one turns up unemployed is equal to the number of ways to pick one unemployed person from 20, times the number of ways to pick four unemployed people out of 30.

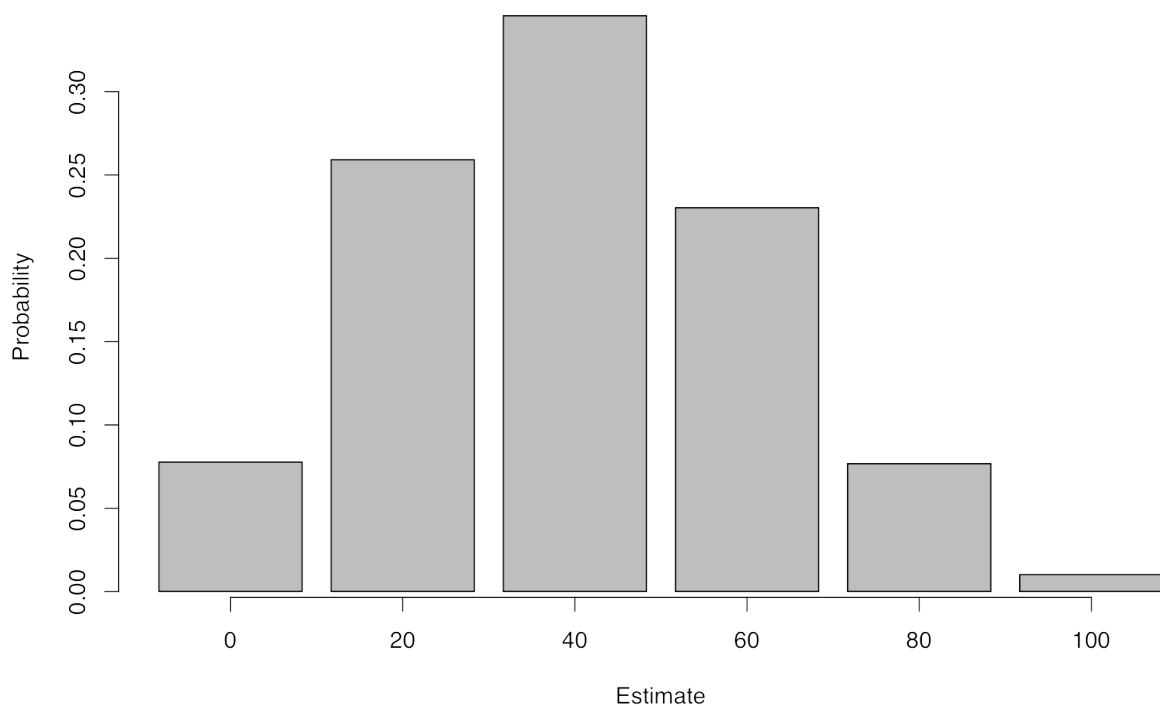
This is written  $\binom{20}{1} \binom{30}{4}$  using the standard notation for "choose." Some readers will recognize a similar<sup>ix</sup> term in the *binomial distribution* function  $B(50,0.4)$ , the formula developed by Bernoulli some time in the 1680s.

This formula makes it possible to tally the number of ways to get a sample with any particular number of unemployed people. Dividing the number of possible samples for each level of unemployment by the total of 2,118,760 possible samples gives the probability of seeing each possible unemployment estimate.

Estimated Unemployment	No. Samples	Probability of Getting This Answer
0%	142,506	0.07
20%	548,100	0.26
40%	771,400	0.36
60%	495,900	0.23
80%	145,350	0.07
100%	15,504	0.01

To make this easier to see we can plot the figures like so:

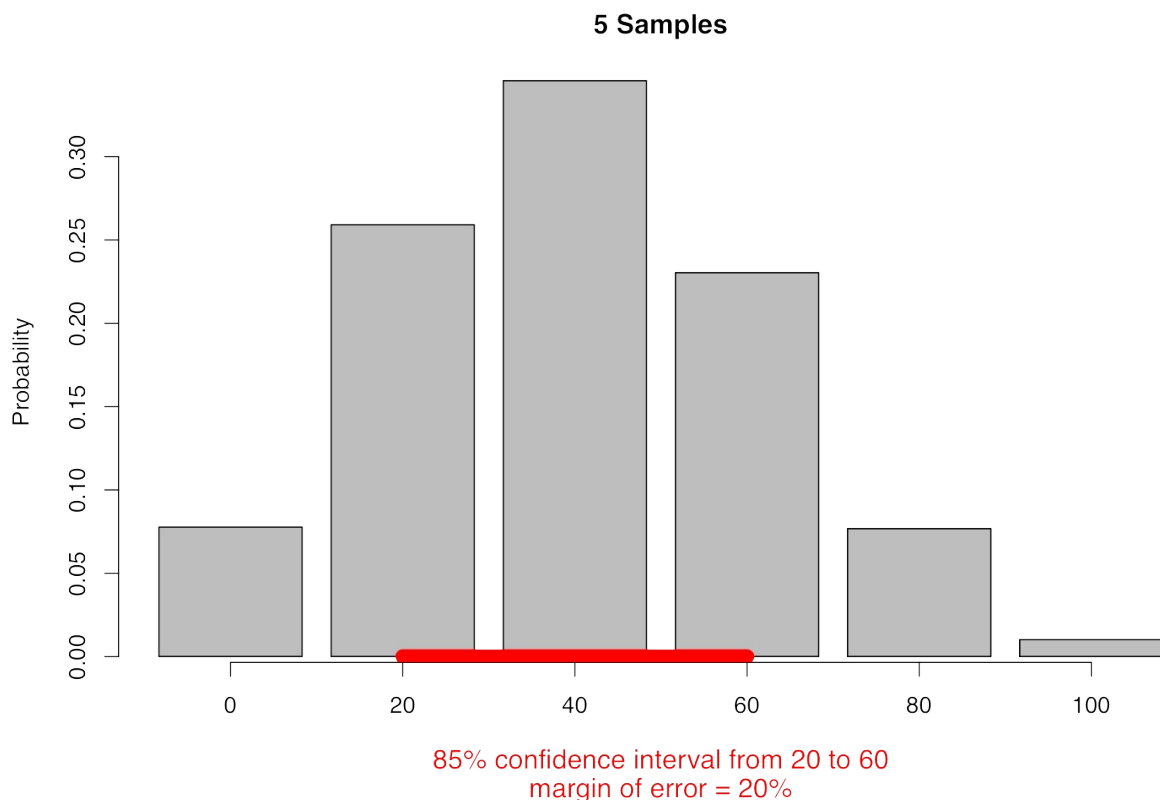




This chart shows a *sampling distribution*, meaning that we would expect to see each answer in these proportions if we repeated the random sampling process many times. As we had hoped, answers closer to the truth occur more often than those further away, and the most common answer is the correct one. There's a probability of 0.36, or a 36 percent chance, that we'll end up with exactly the right answer from our little survey.

This distribution tells us everything we can know about the possible error in our sample value. But we'll often want a more understandable summary, and one way of summarizing an error distribution is to say how often we'll get within a certain distance of the correct answer. Let's say we want to know how often we can expect to get either the true answer of 40%, or the closest incorrect answers of 20% and 60%. This requires adding up the probabilities that we get 20%, 40%, or 60%, which corresponds to seeing one, two, or three unemployed people our sample. There's a probability of  $0.26 + 0.36 + 0.23 = 0.85$  that we'll see any of these three answers.

Among the 2,118,760 different samples of five that we could draw from our population of 50 people, we find that 1,815,400 or 85 percent of them contain one, two, or three unemployed people. Put another way, 85 percent of all samples contain between 20% and 60% unemployed. <sup>X</sup> is known as an 85-percent *confidence interval*. Because this interval covers a 40% range, and our best estimate is right in the middle, we say that the estimate has a margin of error of 20%. The *margin of error* is always half of the width of the confidence interval.



We need one more step. So far we've been talking about the possible samples we might get for a given true unemployment rate of 40%, and how often we'll end up with each estimated number. In reality we never get to know the true unemployment rate! We only ever get one sample, and this gives us only a single error-prone estimate. Instead of "how often is the estimate within the margin of error of the true value," the question we really need to ask is "how often will the true value be within the margin of error of the estimate?"

To do this, we start with the estimated unemployment rate, that is, the rate of unemployment in the actual sample we have. We assume that this is the true rate and construct a margin of error using the process above. If the estimate is within 20% of the true value, then it follows that the true value is within 20% of the estimate. This isn't perfectly accurate, because the margin of error varies in width depending on the true value, so our estimated margin of error won't be quite right if the estimate isn't quite right. You can work out more precise formulas, but this simple method of substituting the estimate for the true value gives a close approximation for practical survey sizes, and it's widely used in practice.

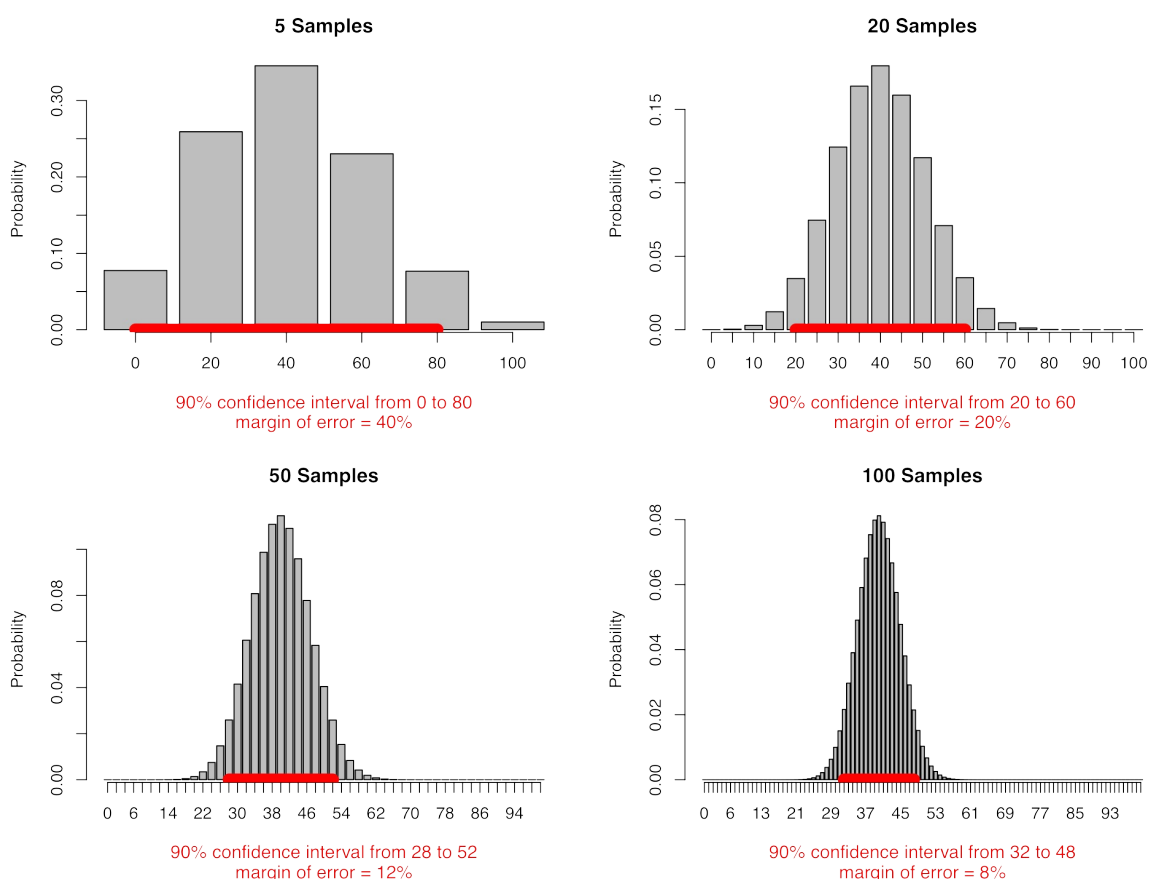
And that's it. We've now calculated the margin of error on our unemployment estimate. There are many different ways of phrasing our result, which all mean the same thing.

- The 85-percent confidence interval is 20% to 60%
- The answer is 40% with a margin of error of 20%, 17 times out of 20.
- We are 85 percent certain that the true answer is between 20% and 60%

- The answer is 40% ± 20% at 85 percent confidence.

Notice that we always use two values to measure the uncertainty: a margin of error and the probability that the true answer falls within that margin of error.<sup>xi</sup> of error, in this case 20% to 60%, is called the 85-percent *confidence interval*. The 85 percent figure itself is called the *confidence level*. Whatever language we use, we have quantified the error in our survey in two values: a range of error and how often you'll see that something within that range.

If 40% ± 20% at an 85-percent confidence level is a precise enough answer, you've reduced your work by a factor of 10 by asking only five out of 50 people. If it's not precise enough, you can sample more people. To compare the error distributions of different numbers of samples, it helps to hold the confidence level constant. The Bureau of Labor Statistics reports the margin of error on unemployment figures at the 90-percent level, so we will too. We'll also do the calculations as if we're sampling from a real country's population, which is much larger than 50.



The accuracy gets better as you ask more people. As the number of samples gets larger—we're up to 100 in the last picture above—the margin of error gets narrower (for a particular confidence level) and the distribution of possible answers rapidly approaches the classic bell-shaped curve, the *normal* distribution. Even better, for large samples the error caused by sampling depends primarily the sample size, not the population size. This means that

estimating the opinions of a hundred million people takes barely more work than estimating the opinions of one million. By the time you survey 1,000 people, the margin of error is down to 3% at the 90-percent confidence level.

This is how we know the error in our monthly unemployment estimates. The Current Population Survey samples 150,000 people out of 300,000,000. The Bureau of Labor statistics has run the math and worked out that it'll get within 300,000 of the true unemployment rate 90 percent of the time, which corresponds to 0.2% difference in the national unemployment rate.<sup>19</sup> The 300,000 is the margin of error and the 90 percent is the confidence level.

If a 90-percent confidence interval sounds like a 10 percent chance of disaster, we can trade off between the estimated error and the risk of falling outside of that error: it's equally true to say that 99 percent of the time the unemployment figures will be accurate to within  $\pm 0.3\%$ . This is the same thing, reported differently; we're just widening the red line on the above charts until it covers 99 percent of the possible outcomes.

There is an intricate bargain being struck here. In exchange for a little fuzziness (the margin of error) and a little risk (the confidence level) we've reduced our work to calculate the unemployment rate by 2,000 times. This remains astonishing to me. It's beautiful and non-obvious and took millennia for humanity to see it.

## The Problem of Measurement Error

In practice, nothing can be measured perfectly.

A random sample has a margin of error due to sampling, but *every* quantification has error for one reason or another. The length of a table cannot be measured much finer than the tick marks on whatever ruler you use, and the ruler itself was created with finite precision. Every physical sensor has noise, limited resolution, calibration problems, and other unaccounted variations. Humans are never completely consistent in their categorizations, and the world is filled with special cases. And I've never seen a database that didn't have a certain fraction of corrupted or missing or simply nonsensical entries, the result of glitches in increasingly complex data-generation workflows.

Error creeps in, and the data never quite matches the description on the box. Anyone who works with data has had this beaten into them by experience.

Even simple counts break down when you have to count a lot of things. We've all sensed that large population figures are somewhat fictitious. Are there really 536,348 people in your hometown, as the number on the "Welcome To ..." sign suggests? If the sign said 540,000, we would know to treat it as a rough figure, yet far too often we're willing to imagine that every last digit is accurate.

There are analogous difficulties with counting the number of people at a protest, the number of intravenous drug users in a city, or the number of stars in the galaxy. Even counting the number of distinct names in a large database can require complex estimation algorithms, given the constraints of distributed storage and finite memory.<sup>20</sup> Large counts are usually estimates, which differ from the true value by some amount.

But we gain hugely if we can say something about the accuracy of our data. Our answer to "how long is the table?" might be "52 inches, to the nearest eighth of an inch."

Reliable data includes measures of error: *how much* the reported information is expected to differ from the reality it represents. There are many standard ways to report the accuracy of different kinds of data. Figures might be "accurate to the nearest quarter pound" or use more technical notation like  $\pm$  and ideas like "standard error" and "confidence interval." For a large database you could report or estimate the number of bad entries. The modern census has a second wave to estimate coverage and therefore error. In many fields it's considered shoddy work to report a figure without giving some idea of the accuracy. Maybe we should say the same for journalism.

The idea of measurement error is the idea of quantified uncertainty. This is one of the tremendous achievements of modern thought—the recognition that knowing *how much* we don't know has great value. Not all data comes with measurement errors attached. Sometimes you have to read the fine print to find out, or call someone and ask. But if you do not know and cannot reasonably guess the sources and magnitudes of possible error, then you don't really know what the data means.

## Quantification Is Representation

The world is very rich and complex. Doesn't trying to reduce it to data lose something vital? Of course!

All quantification throws out information. It has to. That's the point of abstraction: to strip away enough detail that it's possible to use powerful general-purpose reasoning tools. Most things are thrown out when you go from three actual apples to "three apples" recorded in a database. We don't know anything about the color and size of the apples, or why they are there, and maybe one of them is half rotten. If we choose "apple" as our sole unit of symbolic representation, we will be blind to everything else.

But in journalism we throw out information all the time when we select whom we talk to, what we include and exclude in our story, and what we choose to write about at all. Quantification represents the world through the systematic creation of data, a limited but powerful way to gather and summarize information.

Fortunately, quantification is neither mysterious nor fixed by nature. Quantification is always a designed process. If there is some reasonable way to quantify what we care about, a marvelous universe of analysis, representation, and prediction techniques open up to us.

Counting is limited, but there are many things that are best known by counting.

# Analysis

*It may well be that several explanations remain, in which case one tries test after test until one or other of them has a convincing amount of support.* - Sherlock Holmes<sup>21</sup>

It's been said that data speaks for itself. This is nonsense.

It's true that going and looking usually beats sitting and thinking. That's the core idea of empiricism and the point of collecting data. And it's true that data can be revealing and insightful. Sometimes you look at a graph and say "aha!" and feel you understand the world a little better. In that moment there is the sensation that the data is speaking, that it tells a clear story.

But the data didn't tell a story, you did. You saw a story that connects the data to the world. Are you right? Ideally, your story is thoughtfully corroborated by many sources. But if you're going to use data as evidence, you have to understand what it does and doesn't say.

This chapter is about how to draw true meanings from true data. There are mathematical rules which say that two plus two never equals five. There are formulas that encapsulate the logic of working with chance and cause. There are basic principles of investigation, such as testing your guesses. And there are fundamental limitations to knowledge, the cases where we must admit we can't know the answer, at least not with the data we have.

This doesn't mean there's a single right answer in every case. All data analysis is really data *interpretation*, and relies on combining data with something else, such as previously known facts or cultural knowledge. Data, on its own, has no meaning at all. Imagine a spreadsheet with no column names. It would just be numbers, indecipherable and useless.

	A	B	C	D	E
1	Sample Data				
2		Q1	Q2	Q3	Q4
3	Revenue	3700	4142	4099	5006
4	Costs Of Goods Sold	-1877	-1748	-1850	-1921
5	Gross Profit	2123	2396	2419	3077
6	Rent	-440	-440	-440	-440
7	Electricity	-212	-240	-242	-308
8	Other Operating Expenditure	-770	-790	-745	-977
9	Total Operating Expenditure	-1422	-1470	-1427	-1725
10	EBITDA	701	926	992	1352



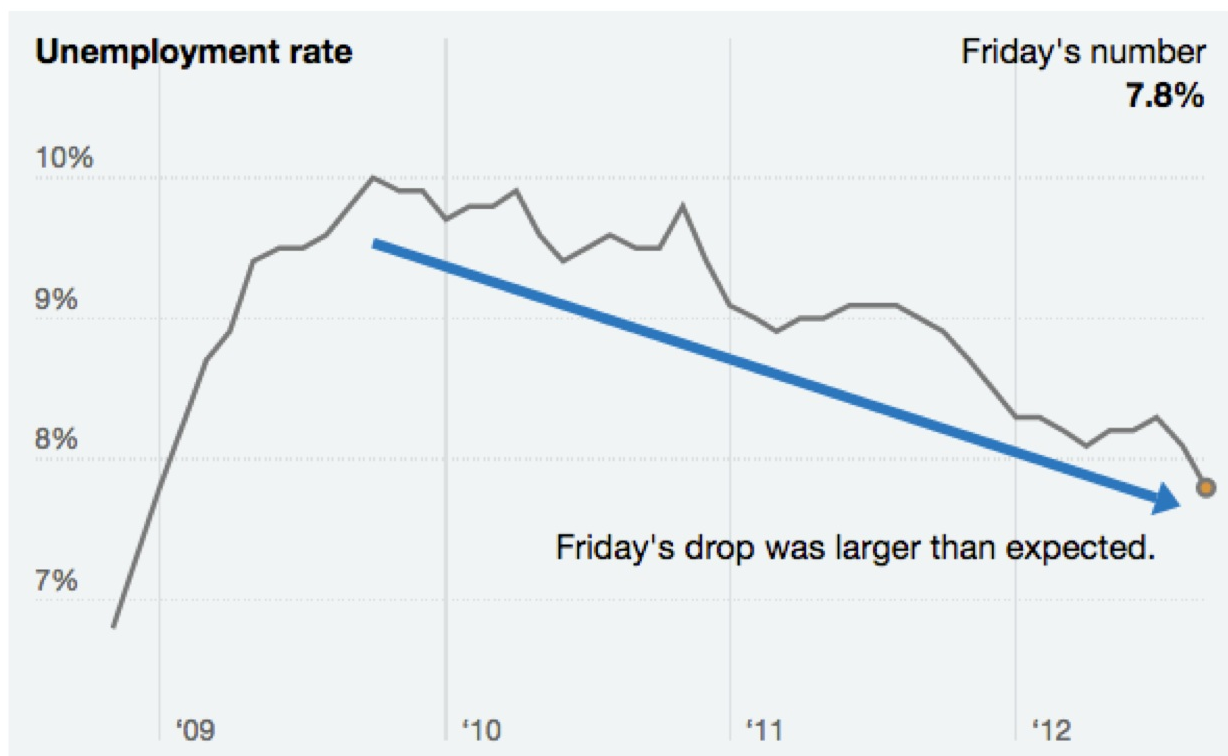
**Data + Context = Meaning**



The necessary context enters in many different ways. Data can't be understood without knowledge of the quantification process that created it. Statistical work usually requires assumptions tied to common knowledge: total kale consumption can't be more than a small fraction of total food consumption, and lower cancer rates are better. But the culture and the journalist are also part of the context that creates meaning. Every society has particular worries that shape what is newsworthy, while individual journalists have specific beats and interests. Actually the context comes before the data; it tells us what data is relevant, even what questions are relevant.

Context is where subjectivity enters into data interpretation. *The New York Times* illustrated this with two different interpretations of the same unemployment data, describing how a Democrat and a Republican might see things.

*The rate has fallen more than 2 points since its recent peak.*

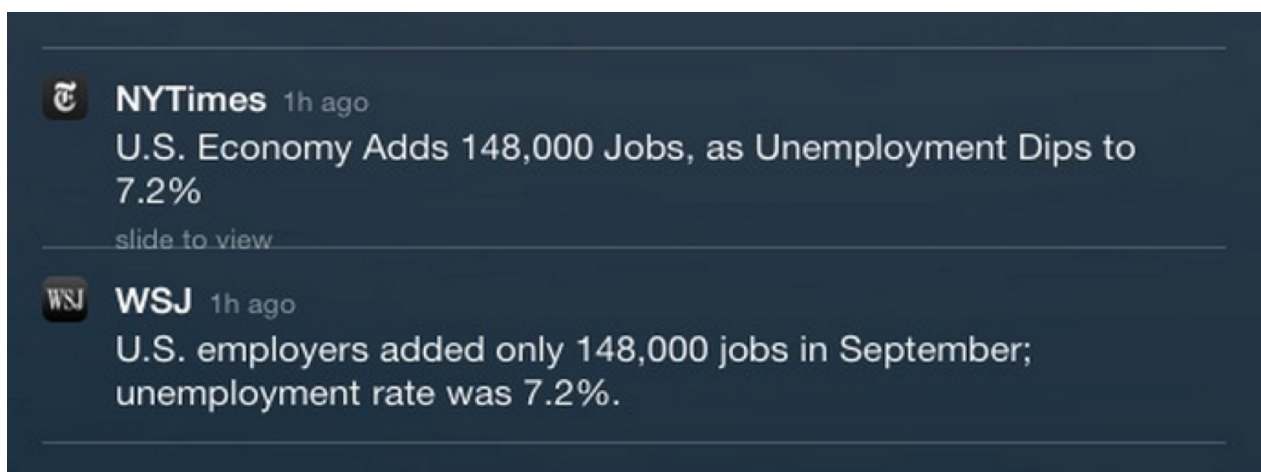


*The rate was above 8 percent for 43 months.*



How Democrats and Republicans might interpret the same unemployment data in different ways.<sup>22</sup>

But it's not just politicians who have different perspectives. Journalists can and do disagree on the interpretation of a single number.



Headlines on October 22, 2013.<sup>23</sup>

Both headlines are perfectly true. The difference between them is down to whether 148,000 merits “only”—is it a big or a small number? This could also be a matter of expectations: perhaps The Wall Street Journal was hoping to see a larger increase in jobs.

This subjectivity may seem disheartening. In the sciences “subjective” is sometimes used as an insult. Subjective things are personal, dependent on who is speaking, maybe a matter of taste. Wasn't data supposed to be objective? Wasn't it supposed to avoid the arbitrariness of opinion and bring us closer to the truth?

Data interpretation may not be mathematical logic, but neither is it nihilist. Our interpretations must be faithful to reality. Out there in the world a policy changed crime rates, or it didn't. The wage gap is some specific level and no other. Careful measurements show climate change is driven by human activity through particular mechanisms, or they don't. All of these are quantitative statements that involve quantification choices—sometimes controversial choices. But once you pick a counting method, reality will see that you end up with a particular number, which is of course the point of counting. Just like a scientist, a journalist can't make up data, ignore evidence, or condone logical fallacies. It's equally important to know when you don't know, when you *can't* answer the question from available data.

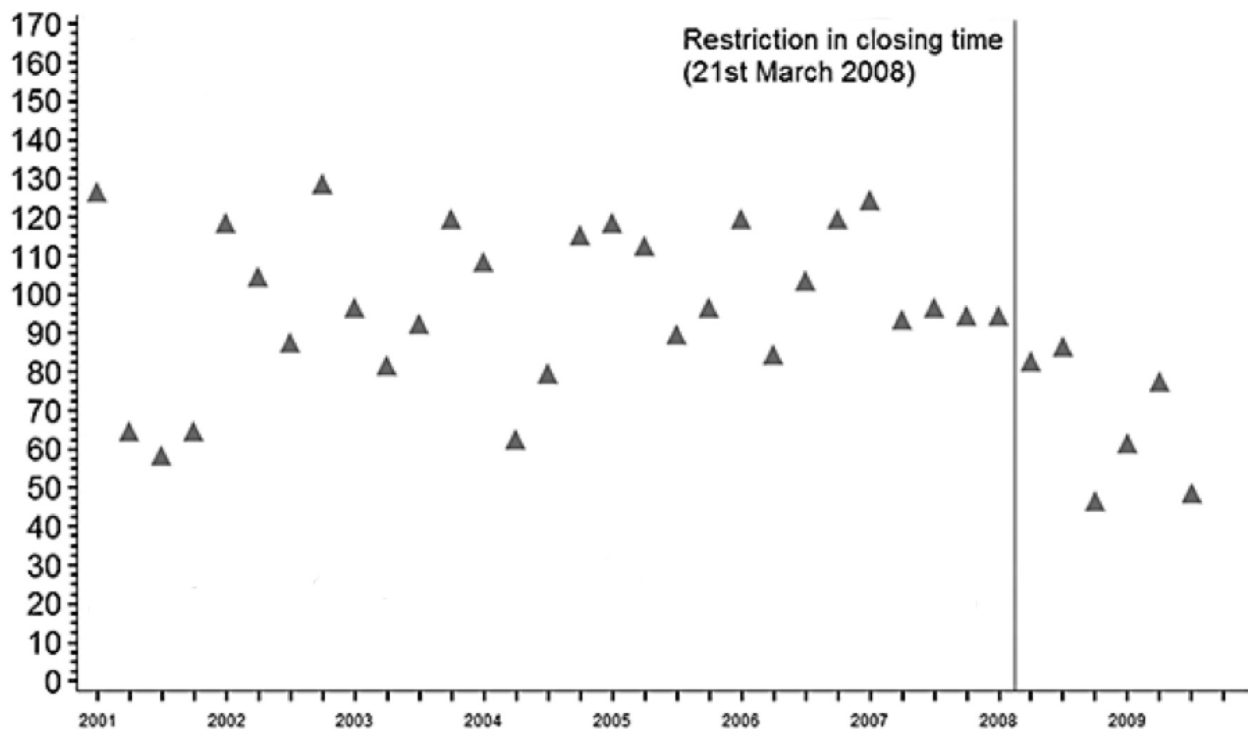
Yet the constraints of truth leave a very wide space for interpretation. There are many stories you could write from the same set of facts, or you could decide that entirely different facts are relevant. Subjectivity is at the core of journalism, because there is no objective theory that tells us which true stories are the best. But “subjective” doesn't necessarily mean “personal.” Culture is widely shared and people live in networks, and journalism requires a broad dose of societal knowledge. Journalists especially need to understand the common knowledge and values of the audience—even if just to challenge them. That audience is never uniform, and different people will have different concerns, experiences, and perspectives. Every time you ask yourself “what is the story here?” you are bringing the audience into your work.

Finding a story in the data will always be an act of cultural creation. But those stories must still be true! So the rest of this chapter is an introduction to three big ideas that can help draw truth from data. The first is the effect of chance, randomness, or noise, which can obscure the real relation between variables or create the appearance of a connection where none exists. The second is the nature of cause, and the situations where we can and can't ascribe cause from the data. Above all is the idea of considering multiple explanations for the same data, rather than just accepting the first explanation that makes sense.

My goal is to give you the higher-level logic of the whole process of statistical analysis. For any particular problem you will need specific technical tools, but those choices must be guided by a larger framework.

## Did the Policy Work?

In 2008 the Australian city of New South Wales had had enough of drunken assaults. The courts imposed an earlier closing time on bars in the central business district: No alcohol after 3 a.m. Now, 18 months later, you have been asked to write a story about whether or not this policy change worked. Here's the data:



*Number of nighttime assaults recorded by police in each quarter in the central business district (CBD) of New South Wales, where closing time was restricted to 3 a.m. Adapted from Kypri, Jones, McElduff and Barker, 2010.<sup>24</sup>*

Our very first questions have to be about the source of the data, the quantification process. Who recorded this and how? Of course the police knew that there was a new closing time being tested—did this influence them to count differently? Even a true reduction in assaults doesn't necessarily mean this is a good policy. Maybe there was another way to reduce violence without cutting the evening short, or maybe there was a way to reduce violence much more.

The first step in data analysis is seeing the frame: the assumptions about how the data was collected and what it means.

But let's assume all of those questions have been asked, and we're down to the question of whether the policy caused a drop in assaults. In principle, there is a correct answer. Out there, in the world, the earlier closing time had some effect on the number of nighttime assaults, something between "nothing at all" to perhaps "reduced by half." Our task is to estimate this effect quantitatively as precisely as possible (and no more precisely than that).

This data is about as clear as you're ever likely to see outside of a textbook. We have about seven years of quarterly data for the number of nighttime assaults in the central district before the new closing time went into effect, and 18 months of data after. After the policy change the average number of incidents is a lot lower, a drop from something like 100-ish per quarter to 60-ish per quarter.

So the policy seems to have worked. But let's spell out the logic of what we're saying here. If you can't express the core of your analysis in plain, non-technical language, you probably don't understand what you're doing. Our argument is:

1. The range of the number of incidents decreased in early 2008.
2. The earlier closing time went into effect around the same time.
3. Therefore, the earlier closing time caused the number of incidents to decrease.

Are we right? There's no necessary reason that the drop in assaults was *caused* by the earlier closing time. The evidence we have is circumstantial, and any other story we could make up to explain the data *might* turn out to be true. That's the core message of this chapter, and the key skill in being right: Consider other explanations.

There are common alternative explanations that are always worth considering.

First, chance. Sheer luck could be fooling us. The actual number of assaults per quarter is shaped by circumstantial factors that we can't hope to know. Who can say why someone threw a punch, or didn't? And we have only six data points from after the new policy went into effect—could we just be seeing a lucky roll of the die?

Second, correlation. The decrease could be related to the earlier closing time without being caused by it. Perhaps the police stepped up patrols to enforce the new law, and it's this increased presence that is reducing crime, not the new closing time itself.

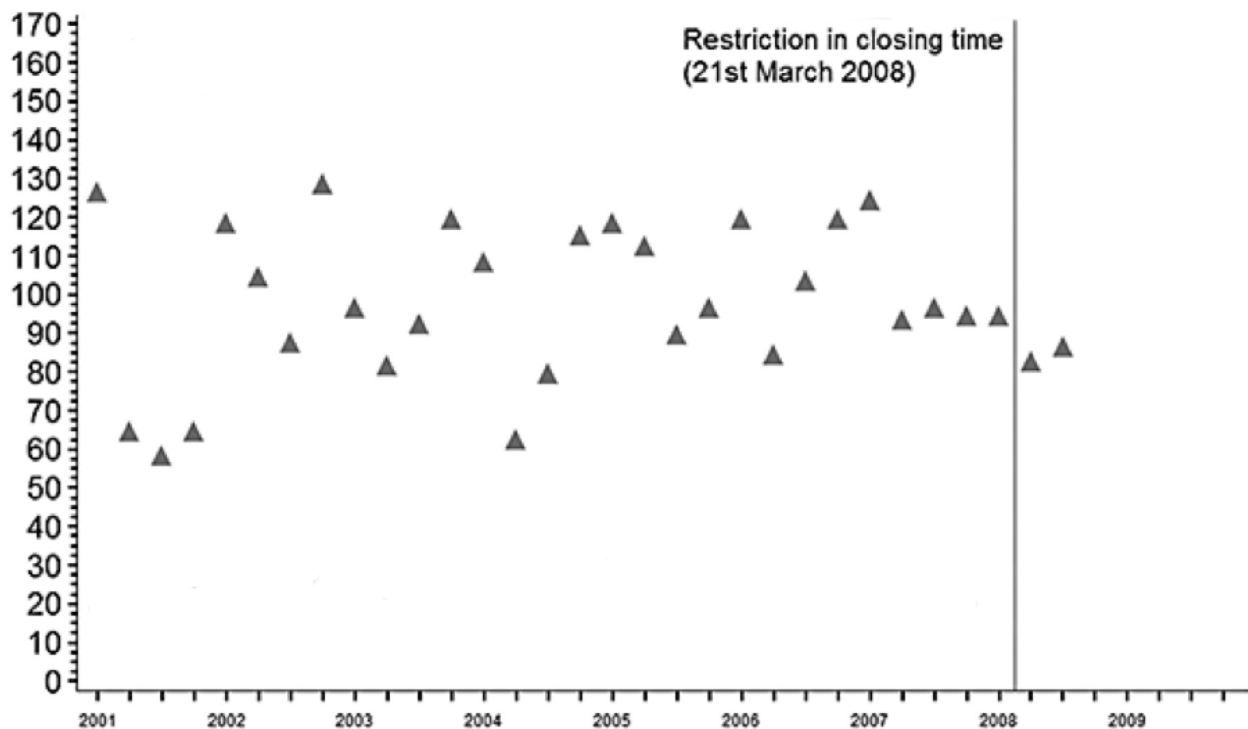
Third, everything else. The change could be caused by something that has never occurred to us. Maybe there was a change in some other sort of policy that has a large effect on nightlife. Maybe crime was falling all over the country at the same time.

We'll tackle these one at a time. To get there, we need to tour through some of the most fundamental and profound ideas of statistical analysis.

## Accounting for Chance

it's very tempting to interpret something as meaningful when it could just as easily be a coincidence—especially if it makes a good story. But dumb luck is always in the running as an explanation for your data. To try to untangle chance from other factors, we can estimate the probability of sheer coincidence.

Our nighttime assaults data shows generous variation. Before the change in closing hours the number of assaults ranged from 60-ish to 130-ish. We say this variation is random, meaning that we can't ever hope to know the circumstances that cause a particular fight on a particular night, and it is precisely this randomness that complicates our analysis.<sup>xii</sup> The less data you have, the more chance is a factor and the easier it is to be fooled. Suppose we only had two quarters of data after the change:



*Number of nighttime assaults, with only two data points after closing time was restricted to 3 a.m. Adapted from Kypri et al.<sup>25</sup>*

If you looked at just this data, you might conclude that the new closing time had no effect. The new points are pretty much in line with the data from the previous four quarters. If anything, it looks like there was a downward shift in the number of assaults a year before the policy ever went into effect! But having seen the additional data, we know that the two points here are at the high end of a new lower range. It's just chance that makes this truncated data look like nothing happened.

If we can be fooled by two chance data points, can we be fooled by six? Certainly, but less probably. How much less?

It takes a while to build up an intuition about the effects of chance. From working with data and models, you eventually get a sense of what randomness looks like, and therefore what it doesn't look like and how much data you need to feel sure about your conclusions. It's well worth getting this sense in your bones. But the great advantage of statistical theory is the ability to quantify chance. "What are the odds that it's just a coincidence?" is not a rhetorical question. It asks for a numeric answer.

## Counting Possible Worlds

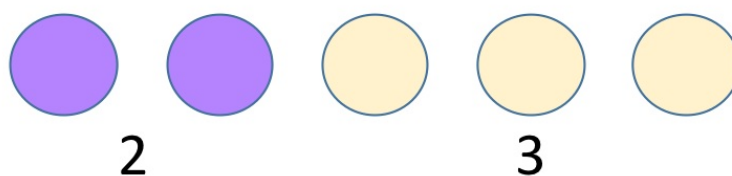
You probably use words like “odds,” “chance,” “frequency,” and “probability” all the time to refer to uncertain events. But before we can go any further we need to get precise about what these words mean. You have to get the basics right or smart people in your audience will make fun of you, and besides you won't be able to calculate anything correctly.

These simple ideas are no less profound for being old and really only emerged in the late 1600s.<sup>xiii</sup> Even if you've been through this before, perhaps I can offer a new perspective. Statistics counts possible worlds.

Probability is a way of reasoning about events that we can't observe. Maybe we can't see what's happening because of practical problems: what's the temperature at the center of the sun? But quite commonly, we will use probability to talk about potential worlds: what would happen if we choose this policy?<sup>xiv</sup> The central insight of probability is that in many of these situations you know more than nothing.

Perhaps you don't don't know what the next roll of the die will be, but you do know that all possibilities will occur in equal proportions. Or you might know that your friend usually orders a blueberry cheesecake at your weekly dinner date, and less commonly the lemon tart. You can use numbers to express these ideas. A probability of 0 means “impossible” while a probability of 1 means “certain,” and all probabilities have to add up 1.

Probabilities are like a percentage in that they are proportions, not counts, and when someone says “percentage chance” they usually mean probability times 100. But it's often more intuitive to think about probabilities as *frequencies*, actual counts of different outcomes. Suppose that over the next five dinners with your friend you would expect her to order two blueberry cheesecakes and three lemon tarts. This hasn't actually happened yet so we're not counting actual deserts, but rather the deserts we expect; probability is a language for talking about our *uncertainty*.



The counts here are frequencies. Probabilities are just the ratio of one type of event to all events.



$$\text{Probability}(\text{purple circle}) = \frac{\text{2 purple circles}}{\text{2 purple circles + 3 yellow circles}} = 2/5$$

The probability that something happens is usually written  $p(\textit{something})$ . In this case  $p(\textit{cake}) = 0.4$ , but like a variable in an equation, you may or may not know the value of your  $p(\textit{something})$ . It may stand in for a number that someone has previously measured or computed, or it may be what you're trying to work out.

The *odds* are a slightly different way of talking about the same proportion.

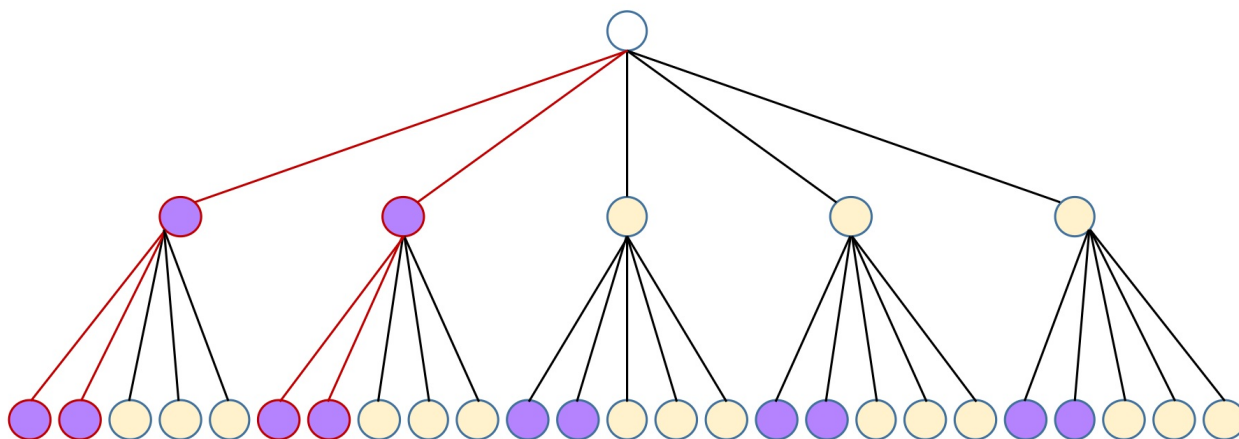
$$\text{Odds}(\text{purple circle}) = \frac{\text{2 purple circles}}{\text{3 yellow circles}} = 2/3$$

The odds are defined as the number of events we are counting divided by the number we are not counting. In gambling the odds are the number of times you win divided by the number of times you don't. The odds of cake here are 2/3 or 0.66, but we usually report odds by giving the numerator and the denominator separately: the odds are 2 to 3. You can convert odds to probability by dividing the first number by the sum of the two: 2 to 3 odds is a probability of  $2 / (2+3)$ . Odds of 1 to 1 mean a probability of  $1 / (1+1) = 1/2$ , or a 50/50 chance.

Although "odds" and "probability" are both numeric measurements of chance, they are different formulas and if you confuse them you will get the wrong answer. Don't be that journalist. (You're also welcome to correct people when they use the wrong words, but remember: pedants die alone.)

We can do some nifty things with simple probabilities. How many cakes do you expect your friend to order over the next 20 dinners? This is just  $p(\textit{cake}) \times 20 = 0.4 \times 20 = 8$ . You can think of 0.4 as the average number of cakes she orders per dinner. Of course there is randomness here; she actually orders either zero or one cakes each time, and over the course of 20 dinners she might order 7 or 9 or 17 cakes, but 8 will be the most common number. (Because there are two possible desert choices, you get a *binomial* distribution just like the sampling distribution from the last chapter.)

Quite often we will need to count how frequently multiple events occur together. What is the probability that your friend orders cheesecake at the next two dinners? Let's draw every possible combination of her first and second dessert orders.



For her first dinner she orders cake 2 out of 5 times. After each of those, she orders cake again 2 out of 5 times. Hence there are  $2 \times 2 = 4$  possible worlds where you get two cake orders in a row. Since there are 25 possibilities in total, the probability is  $4/25$  or 0.16.

Or, we could just multiply  $p(\text{cake}) \times p(\text{cake}) = 0.4 \times 0.4 = 0.16$ . The definition of probability divides out the total number of cases so that probabilities are always between 0 and 1, which lets us avoid the tedious bookkeeping of counting cases directly when all we want is the final proportion. Multiplication is how you work out the probability that event A *and* event B both happen when the events in question are *independent*, that is, one doesn't affect the other. Whether or not this is true is a question your data cannot answer. A coin doesn't care if it came up heads or tails last time, but maybe your friend will get tired of too many cakes in a row.

We can apply the multiplication rule to our assaults data. Suppose we can work out the probability that we'll see a quarter with 80 or fewer assaults just by chance, even if the earlier closing time did nothing. Call this  $p(\text{low})$ . Then the probability that we'll see two low quarters in a row is  $p(\text{low}) \times p(\text{low})$ , the probability of seeing three low quarters in a row is  $p(\text{low}) \times p(\text{low}) \times p(\text{low})$ , and so on.

In practice you don't work out probabilities by drawing trees, just as you don't work out the margin of error by drawing pictures of samples. Still, I love thinking in terms of trees of possibilities because it makes plain what we are doing with probability arithmetic. Each branch is a possible course through history, and we are assigning probabilities by counting the branches of different types. All of statistics is based on the idea of counting possibilities.

## Arguing From the Odds

We can use the logic of counting cases to work out the probability of an unlikely event happening by chance. In the winter of 1976 the United States embarked on a nationwide flu vaccination program, responding to fears of an H1N1 virus epidemic (a.k.a. swine flu). Millions of people lined up across the country to get vaccinated. But some of them got sick after, or even died. *The New York Times* wrote an editorial:

It is disconcerting that three elderly people in one clinic in Pittsburgh, all vaccinated within the same hour, should die within a few hours thereafter. This tragedy could occur by chance, but the fact remains that it is extremely improbable that such a group of deaths should take place in such a peculiar cluster by pure coincidence.<sup>26</sup>

But is it really “extremely improbable?” Nate Silver has estimated the odds:

Although this logic is superficially persuasive, it suffers from a common statistical fallacy. The fallacy is that, although the odds of three particular elderly people dying on the same particular day after having been vaccinated at the same particular clinic are surely fairly long, the odds that some group of three elderly people would die at some clinic on some day are much shorter. > > Assuming that about 40 percent of elderly Americans were vaccinated within the first 11 days of the program, then about 9 million people aged 65 and older would have received the vaccine in early October 1976. Assuming that there were 5,000 clinics nationwide, this would have been 164 vaccinations per clinic per day. A person aged 65 or older has about a 1-in-7,000 chance of dying on any particular day; > the odds of at least three such people dying on the same day from among a group of 164 patients are indeed very long, about 480,000 to one against. However, under our assumptions, there were 55,000 opportunities for this “extremely improbable” event to occur—5,000 clinics, multiplied by 11 days. The odds of this coincidence occurring somewhere in America, therefore, were much shorter—only about 8 to 1 against.<sup>27</sup>

This is a mouthful. It doesn't help that Silver is switching between probabilities (“a 1-in-7000 chance”) and odds (“480,000 to one”). But it's just a bunch of probability arithmetic. The only part that isn't simple multiplication is “the odds of at least three such people dying.” In practice your calculator will have some command to solve these sorts of counting problems. The more fundamental insight is that you can multiply the probability of three people dying on the same day in the same city by the number of opportunities where it *could* happen to work out how often it *should* happen.

To be sure, this can only be a rough estimate; there is a big pile of assumptions here, such as the assumption that death rates don't vary by place and time. But the point of this exercise is not to nail down the decimals. We're asking whether or not we should believe that chance is a good explanation for seeing three post-vaccination deaths in one day, and we only need an order-of-magnitude estimate for that. Rough estimates can be incredibly useful for checking your story, and there's a trove of practical lore devoted to them.<sup>28</sup>

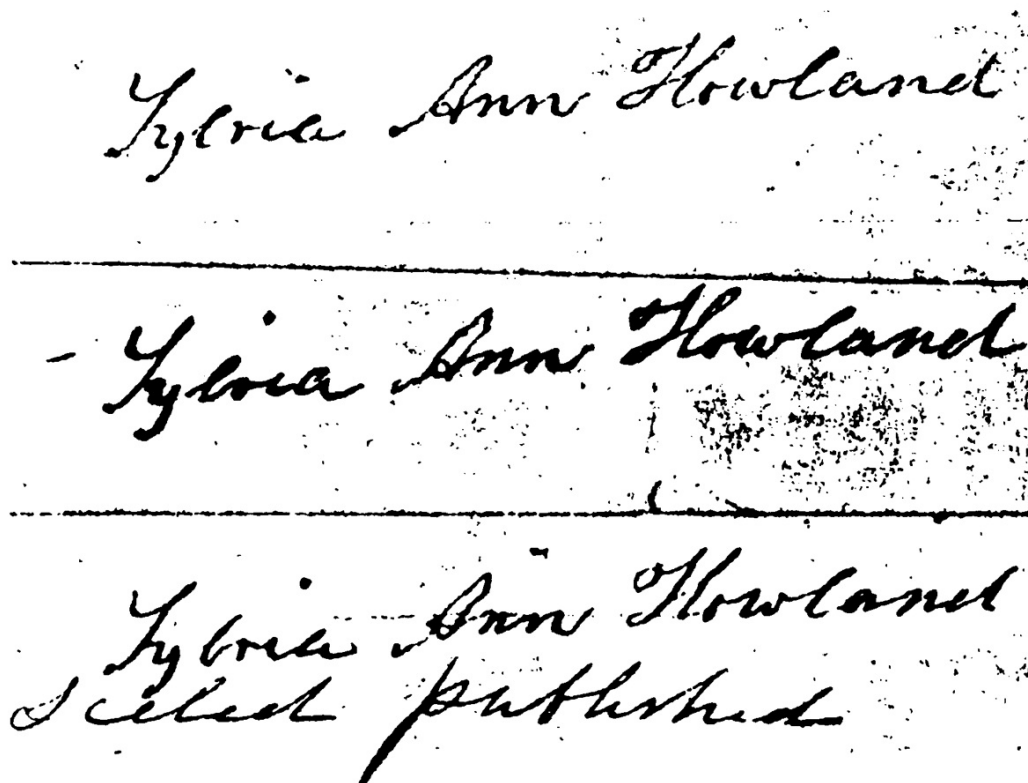
The odds "8 to 1 against" is a probability of 1/9, or an 11 percent chance that we'd see three people from the same clinic die on the same day. Is this particularly long odds? This question is hard to answer on its own.

The less likely it is that something can occur by chance, the more likely it is that something other than chance is the right explanation. This sensible statement is no less profound when you think it through. This idea emerged in the 1600s when the first modern statisticians asked questions about games of chance. If you flip a coin 10 times and get 10 heads, does that mean the coin is rigged or are you just lucky? The less likely it is to get 10 heads in a row from a fair coin, the more likely the coin is a fake. This principle remains fundamental to the disentangling of cause and chance.

Coins and cards are inherently mathematical. Random deaths are a sort of lottery, where you can multiply together the probabilities of the parts. It can be a little harder to see how to calculate the probabilities in more complex cases. The key is to find some way of quantifying the randomness in the problem. One of the earliest and most famous examples of accounting for chance in a sophisticated way concerns a fake signature, millions of dollars, and a vicious feud of the American aristocracy.

In 1865, Sylvia Ann Howland of Massachusetts died and left behind a 2,025,000-dollar estate—that would be about 50 million dollars today. But the will was disputed, there was a lawsuit, and the plaintiff argued that the signature was traced from another document. To support this argument, the mathematician Benjamin Peirce was hired to prove that the original signature could not match the disputed signature so closely purely by chance. The signatures looked like this:

**A. The Unquestioned Original Signature (no. 1) and the Two Disputed Signatures (nos. 10 and 15)**



*A known genuine and two possibly forged signatures in the Howland will case. From Meier and Zabell, 1980.<sup>29</sup>*

To work out the probability of these two signatures matching by chance, Peirce first worked out how often a single stroke would match between two authentic signatures. He collected 42 signatures from other documents, all of them thought to be genuine. Then he instructed his son, Charles Sanders Peirce, to superimpose each of the 861 possible pairs of these 42 signatures and count how many of the 30 downward-moving strokes aligned in position and length. Charles found that the same stroke in two different signatures matched only one-fifth of the time. This is the key step of quantifying random variation, which Peirce did by counting the coincidences between signatures produced in the wild.

But every stroke of every letter matched exactly between the original and disputed signatures. The elder Peirce wanted to show just how unlikely it was that this could happen by chance, so he assumed that every stroke was made independently which allowed him to use the multiplication rule for probabilities. Since there are 30 strokes in the signature and a  $1/5$  chance of any single stroke matching, he argued that the positions of the strokes of two genuine signatures should match by chance only once in  $5 \times 5$  times, that is, once in  $5^{30}$ . This is a fantastically small number, a 0.00000000000000000001 percent chance of a random match. According to this calculation, if you signed your name like Mrs. Howland and did it a billion times you would never see the same signature twice; one in a

billion would be a much healthier 0.0000001 percent chance. A modern analysis which does not assume independence of each stroke gives a probability several orders of magnitude more likely, but still extraordinarily unlikely.<sup>xv</sup>

It seemed much more likely that the signature was forged by Hetty Robinson, Sylvia Ann Howland's niece who was contesting the will. Robinson had access to the original documents and stood to gain millions of dollars by tracing Mrs. Howland's signature on an extra page spelling out favorable revisions.

I admit I'm disappointed that the case was ultimately decided on other grounds, rendering this analytical gem legally irrelevant. But the event was a milestone in the practical use of statistics. Statistics was mostly applied to physics and gambling at that time, never anything as qualitative as a signature. The trick here was to find a useful way of quantifying the variations from case to case. Charles Sanders Peirce went on to become one of the most famous nineteenth-century scientists and philosophers, contributing to the invention of the randomized controlled experiment and the philosophical approach known as pragmatism.<sup>30</sup>

The probability that you would see data like yours purely by chance is known as the p-value in statistics, and there is a popular theory of statistical testing based on it. First, you need to choose a test that defines whether some data is "like yours." Peirce said a pair of signatures is "like" the two signatures on the will if all 30 strokes match. Then imagine producing endless random data, like scribbling out countless signature, or monkeys banging on typewriters. Peirce couldn't get the deceased Howland to write out new pairs of signatures, so he compared all combinations of all existing known genuine signatures. The p-value counts how often this random data passes the test of looking like your data—the data you suspect is not random.

There's a convention of saying that your data is *statistically significant* if  $p < 0.05$ , that is, if there is a 5 percent probability (or less) that you'd see data like yours purely by chance. Scientists have used this 5 percent chance of seeing your data randomly as the minimum reasonable threshold to argue that a particular coincidence is unlikely to be luck, but they much prefer a 1 percent or 0.1 percent threshold for the stronger argument it makes.<sup>31</sup> But be warned: No mathematical procedure can turn uncertainty into truth! We can only find different ways of talking about the strength of the evidence. The right threshold to declare something "significant" depends on how you feel about the relative risks of false negatives and false positives for your particular case, but the 5 percent false positive threshold is a standard definition that helps people communicate the results of their analyses.

Let's use this  $p < 0.05$  standard to help us evaluate whether the 1976 flu vaccine was dangerous. By this convention, an 11 percent chance of seeing three people randomly die on the same day is evidence against a problem with the vaccine; you could say the occurrence of these deaths is not statistically significant. That is, because there is a greater

than 5 percent chance that we'd see data like ours (three people dying) even if the vaccine is fine, it's not a good bet to assume that these deaths were caused by a toxic vaccine. But this does *not* mean there is an 11 percent chance that the vaccine is safe. We haven't yet said anything at all about the vaccine; so far we've only talked about the odds of natural death.

Really the question we need to ask is comparative: Is it more likely that the vaccine is harmful, or that the three deaths were just a fluke? And how much more likely? Is there greater or less than an 11 percent chance the vaccine is toxic and no one noticed during earlier testing? In the case of the Howland will, we found miniscule odds that two signatures could end up identical by accident. But what are the odds that Mrs. Howland's niece forged the will? A more complete theory of statistics tests multiple alternatives.



## Statistical Inference

There is a completely general method of accounting for chance which forms the basis of modern statistical reasoning. *Inference* is the process of combining existing knowledge to get new conclusions, something we do every day. *Statistical inference* adds the element of uncertainty, where both our information and our conclusions have an element of chance.

The propositional logic of the Greeks gave us a template for reasoning when every variable is exactly true or false: “If it rains, the grass will get wet. The grass is not wet. Therefore it did not rain today.” The theory of statistical inference extends this to uncertain information and uncertain answers: “There was a 40 percent chance of rain today. it’s hard to say from just looking out my window, but I’m 70 percent sure the grass is dry. What’s the probability that it rained today?”

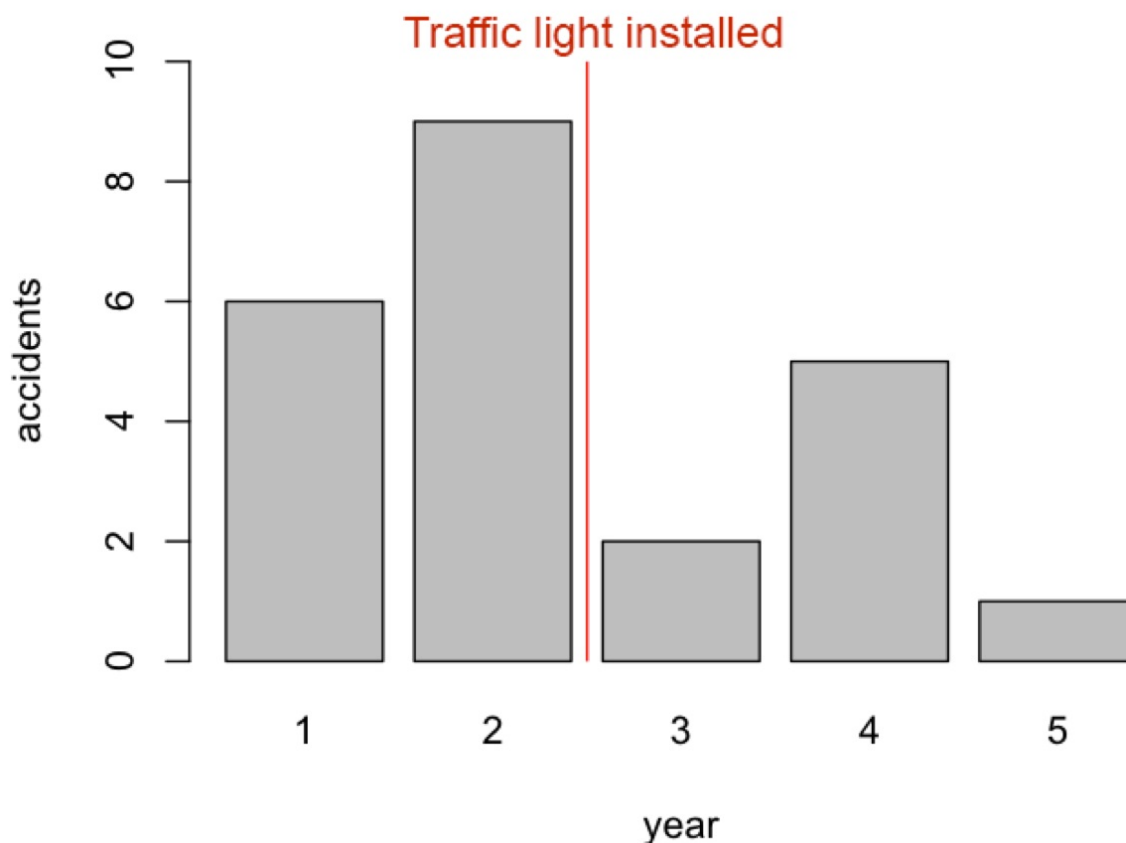
The most comprehensive modern theory is usually called *Bayesian* statistics after its roots in Reverend Bayes’s theorem of 1763. But the practical method was only fully developed in the twentieth century with the advent of modern computing. If you’ve never seen this sort of thing before, it’s unlikely that this little introduction will prepare you to do your own analyses. We can’t cover all of Bayesian statistics in a few pages, and anyway there are books on that.<sup>xvi</sup> walk through a specific Bayesian method, a general way to answer multiple-choice questions when the answer is obscured by randomness. My purpose is to show the basic logic of the process, and to show that this logic is commonsensical and understandable. Don’t let statistics be mysterious to you!

Bayesian statistics works by asking: What hypothetical world is most likely to produce the data we have? And how much more likely is it to do so than the alternatives? The possible “worlds” are captured by statistical *models*, little simulations of hypothetical realities that produce fake data. Then we compare the fake data to the real data to decide which model most closely matches reality.

With the multiple-choice method in this chapter you can answer questions like “how likely is it that the average number of assaults per quarter really decreased after the earlier closing time?” Or “if this poll has Nunez leading Jones by 3 percent but it has a 2 percent margin of error, what are the chances that Nunez is actually the one ahead?” Or “could the twentieth century’s upward global temperature trend be just a fluke, historically speaking?”

We’ll work through a small example that has the same shape as our assaults versus closing time policy question. Suppose there is a dangerous intersection in your city. Not long ago there were nine accidents in one year! But that was before the city installed a traffic light. Since the stoplight was installed there have been many fewer accidents.





Accident data surely involves many seemingly random circumstances. Maybe the weather was bad. Maybe a heartbroken driver was distracted by a song that reminded them of their ex. A butterfly flaps its wings, etc.<sup>xvii</sup> Nonetheless, it is indisputably true that there were fewer accidents after the stoplight was installed.

But did the stoplight actually reduce accidents? We might suspect that a proper stoplight will cut accidents in half, but we have to regard this possibility as a guess, so we say it's a *hypothesis* until we find some way to prove it. We're going to compare the following hypotheses:

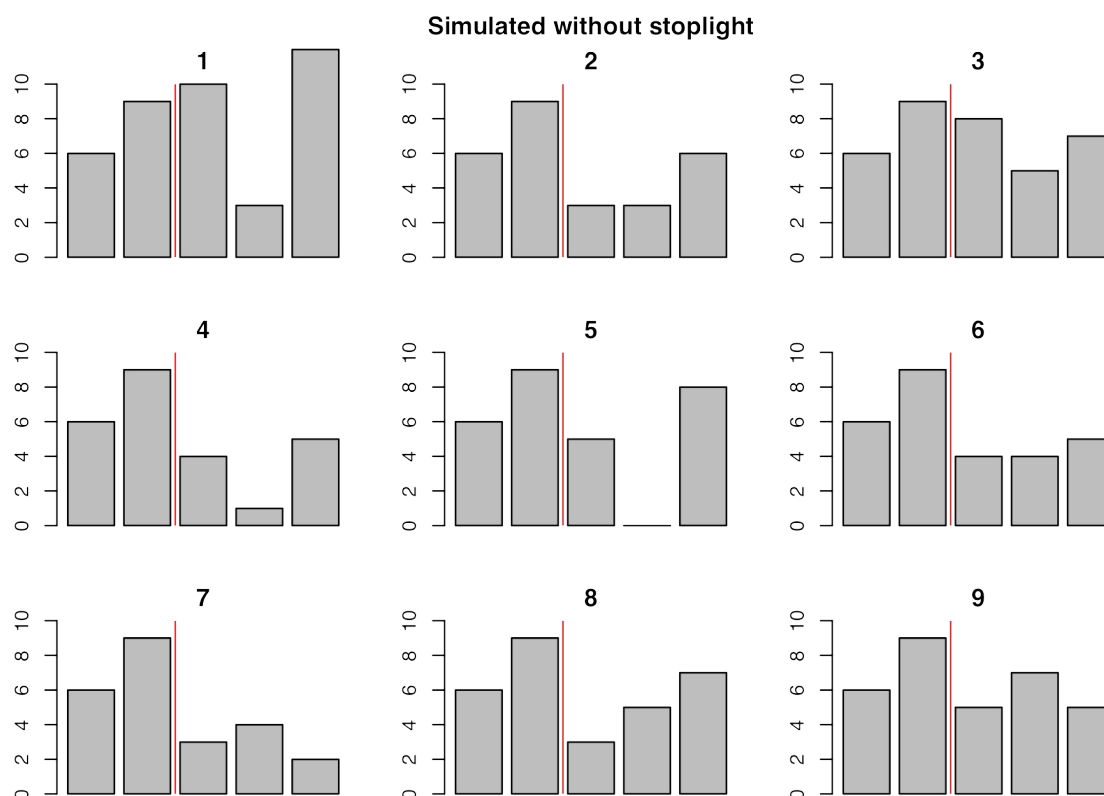
1. The stoplight was effective in reducing accidents by half.
2. The stoplight did nothing, meaning that the observed decline in accidents is just luck.

The next thing we need is a statistical *model* for each hypothesis. A model is a toy version of the world that we use for reasoning. It incorporates all our background knowledge and assumptions, encapsulating whatever we might already know about our problem. Silver used a simple model, based on the odds of any given person dying on any given day, to estimate the odds of three people dying on the same day at any of 5,000 clinics. Peirce created a model based on the stroke positions of 42 signatures that were known to be genuine. A model is by definition a fake. It's not nearly as sophisticated as reality. But it can be useful if it represents reality in the right way. Creating a model is a sort of quantification step, where we encode our beliefs about the world into mathematical language.

For our purposes a model is a way to generate fake data, imagined histories of the world that never occurred. We'll need two assumptions to build a simple model of our intersection. We'll assume that the same number of cars pass each day, and we'll pick the number based on the historical data we have. We'll further assume that there is some percentage chance of each car getting into an accident as it does, and again we'll use historical data, pre-stoplight, to guess at the proper percentage.

With these two numbers in hand you can imagine writing a small piece of code to simulate the intersection. As each simulated car goes into the simulated intersection we can flip a simulated coin to determine whether to count an accident. We calibrate the "coin" so the cars crash at the proper percentage. This is a reasonable model if we are willing to assume that car accidents are independent: there might have been an accident at this intersection a year or an hour ago but that doesn't change the odds that *you* are about to have an accident.<sup>xviii</sup>

By setting up the simulation to produce the same average accident rate as we saw pre-stoplight, we've built a model of the intersection without the stoplight that we hope matches the real world. We can use this model to get a feel for the range of scenarios that chance can produce by running the simulation many times, like this:



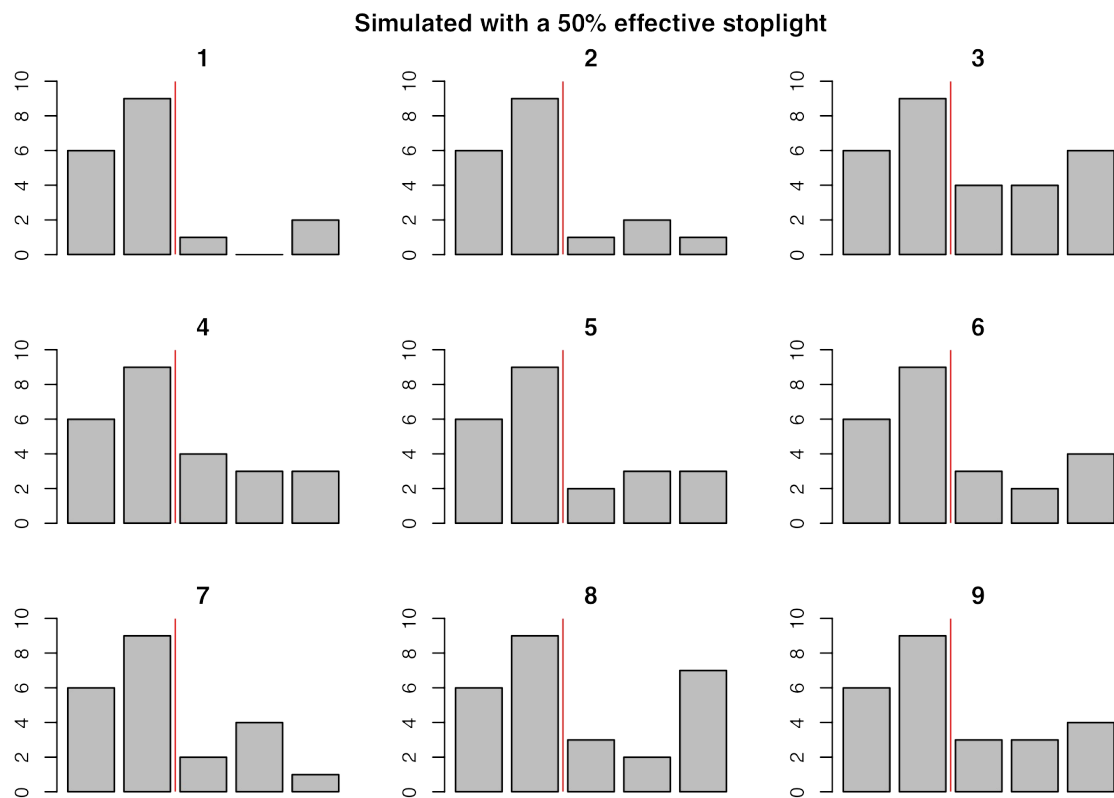
The first two years in each of these charts are just the original data, pre-stoplight. The last three years have been generated by simulation. In some of these alternate histories the number of accidents decreased relative to the pre-stoplight years, and in others the pattern was increasing or mixed, all purely by chance. In order to compare models, we first need to pick a more precise definition of “decline.” So let’s say that the accidents “declined” if all the post-stoplight years show fewer accidents than any of the pre-stoplight years—just like the real data from the actual intersection. This is a somewhat arbitrary criterion, but your choice determines exactly which hypotheses you are testing. Just as our simulation expresses the world in code, our test criterion expresses the hypotheses mathematically. By our chosen test, scenarios 4, 6, and 7 show a decrease in the accident rate. We are counting the branches of a tree of possibilities once more.

The key number is how often we see the effect without the alleged cause, just like the vaccine deaths and Howland will case. None of these alternate histories include a stoplight, yet we see a decline after the second year in 3/9 cases, which is a probability of 0.33. This makes the “chance decline” theory pretty plausible. A probability of 0.33 is a 33 percent chance, which may not seem “high” compared to something that happens 90 percent of the time, but if you’re rolling dice you’re going to see anything that happens 33 percent of the time an awful lot.

This doesn’t make the “chance decline” hypothesis *true*. Or false. It especially does not mean that the chance decline theory has a 33 percent chance of being true. We *assumed* that “chance decline” was true when we constructed the simulation. In the language of *conditional probability*, we have computed  $p(\text{data} \mid \text{hypothesis})$  which is read “the probability of the data given the hypothesis.” What we really want to know is  $p(\text{hypothesis} \mid \text{data})$ , the probability that the hypothesis is true given the data. The distinction is kind of brain bending, I admit, but the key is to keep track of which way the deduction goes.

As we saw in the last section, the more likely it is that your data was produced by chance, the less likely it was produced by something else. But to finish our analysis we need a comparison. We haven’t yet said anything at all about the evidence for the “stoplight worked” theory.

First we need a model of a working stoplight. If we believe that a working stoplight should cut the number of accidents in half in an intersection like this, then we can change our simulation to produce 50 percent fewer accidents. This is an arbitrary number; a more sophisticated analysis would test and compare many possible numerical values for the reduction in accidents. Here’s the result of simulating a 50 percent effective stoplight many times:



Again, each of these charts is a simulated alternate history. The first two years of data on each chart is our real data and the last three years are synthetic. This time the simulation produces half as many accidents on average for the last three years, because that's how effective we believe the stoplight should be. By our criterion that every post-stoplight year should be lower than every pre-stoplight year, there's a reduction in accidents in simulations 1, 2, 4, 5, 6, 7, and 9. This is 7 out of 9 scenarios declining, or a  $7 / 9 = 0.78$  probability that we'd see a decline like the one we actually saw, if the stoplight reduced the overall number of accidents by half.

This is good evidence for the "stoplight cut accidents in half" hypothesis. But the probability of seeing this data by chance is 0.33, which is also pretty good. This is not a situation like Mrs. Howland's will where the odds of one hypothesis were miniscule (identical signature by chance) while the odds of the other hypothesis were good (forged signature to get millions of dollars).

Finally we arrive at a numerical comparison of two hypotheses in the light of chance effects. The key figure is the ratio of the probabilities that each model generates data like the data actually observed. This is called the *likelihood ratio* or *Bayes factor*, and you can think of it as the odds in favor of one model as compared to another. The key idea of comparing multiple models was fleshed out in the early twentieth century by figures such as R. A. Fisher<sup>32</sup> and Harold Jeffreys.<sup>33</sup>

The probability that “stoplight cut accidents in half” could generate our declining data is 0.78 while the probability that “chance decline” accounts for the data is 0.33, so the Bayes factor is  $0.78 / 0.33 = 2.3$ . This means that the odds of the “stoplight worked” model generating the observed data, when compared to the “chance decline” model, are 2.3 to 1 in favor.

This doesn't make the “stoplight cut accidents in half” story true. But it definitely seems more likely.

These 2.3 to 1 odds are middling. Converting the odds to a probability, that's a  $2.3 / (2.3+1) = 70$  percent chance the stoplight worked. That means if you write a story which says it did work, there's a 30 percent chance you're wrong. In other situations you might have a 90 percent or 99 percent or even 99.9 percent chance of guessing correctly. But there can be no fixed scale for evaluating the odds, because it depends on what's at stake. Would 2.3 to 1 odds be good enough for you to run a story that might look naive later? What if that story convinced the city government to spend millions on stoplights that didn't work? What if your story convinced the city government *not* to spend millions on stoplights that did work, and could have saved lives?

Even so, “stoplight worked” is a better story than “chance decline.” A better story than either would be “stoplight probably worked.” Journalists, like most people, tend to be uncomfortable with intermediate probability values. A 0 percent or 100 percent chance is easy to understand. A 50/50 chance is also easy: You know essentially nothing about which alternative is better. It's harder to know what to do with the 70/30 chance of our 2.3 to 1 odds. But if that's your best knowledge, it's what you must say.

In real work we also need to look at more than the data from just one stoplight. We should be talking to other sources, looking at other data sets, collecting all sorts of other information about the problem. Fortunately there is a natural way to incorporate other knowledge in the form of *prior odds*, which you can think of as the odds that the stoplight worked given all other evidence *except* your data. This comes out in the mathematical derivation of the method, which says we need to multiply our Bayes factor of 2.3 to 1 by the prior odds to get a final estimate.

Maybe stoplight effectiveness data from other cities shows that stoplights usually do reduce accidents but seem to fail about a fifth of the time, so you pick your prior odds at 4 to 1. Multiplying by your 2.3 to 1 strengthens your final odds to 9 to 1. The logic here is: stoplights in other cities seem to work, and this one seems to work too, so the totality of evidence is stronger than the data from just this one stoplight.

Or maybe you have talked to an expert who tells you that stoplights usually only work in large and complex highway intersections, not the quiet little residential intersection we're looking at, so you pick prior odds of 1 to 5, which could also be written 0.2 to 1. In this case even our very plausible data can't overwhelm this strong negative evidence, and the final

odds are  $2.3 \times 0.2 = 0.46$  to 1, meaning that it's more than twice as likely that the stoplight didn't work. The logic here is: most stoplights at this kind of intersection don't work, and this undermines the evidence from this one stoplight, which leads us to believe that the observed decline is more likely than not just due to chance.

Multiplying by the prior is mathematically sound, yet it's often unclear how to put probabilities on available evidence. If the mayor of Detroit tells you she swears by stoplights in her city, what does this say about the odds of stoplights working versus not working as a numeric value? There is no escape from judgment. But even very rough estimates may be usefully combined this way. If nothing else, the existence of the prior in statistical formulas helpfully reminds us to consult all other sources!

There is a lot more to say about this method of comparing the likelihood that different models generated your data. The method here only applies to multiple-choice questions, whereas real work often estimates a parameter: how *much* did the stoplight reduce accidents? And we've barely touched on modeling, especially the troubling possibility that all of your models are such poor representations of reality that the calculations are meaningless.<sup>xix</sup> But the fundamental logic of comparing how often different possibilities would produce your observed data carries through to the most complex analyses. I hope this example gives the flavor of how a single unifying framework has been used to solve problems in medicine, cryptography, ballistics, insurance, and just about every other human activity.<sup>34</sup> Bayesian statistics is something remarkable, and I find its wide success incredible, unlikely, and almost shockingly too good to be true. You can always start from the general framework and work your way toward the details of your problem. This is sometimes more work, but it is the antidote to staring at equations and wondering if they apply.

## What Would Have Happened Anyway?

Let's suppose we've ruled out luck as an explanation for our data. Suppose we have inferred that something in the assaults data really did change around the time the new closing-time policy came into effect. Attributing this change to the new closing times is another matter entirely.

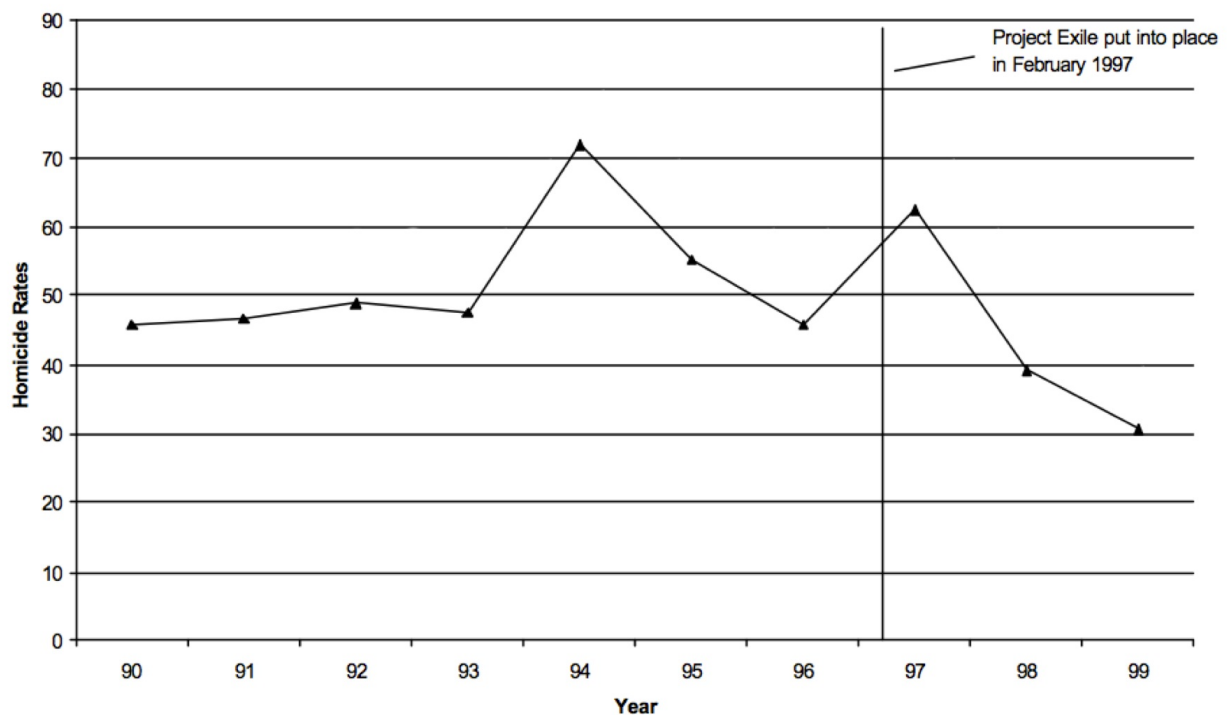
It would be easy to determine the true effects of the new policy if we knew how many assaults we would have seen had the policy never gone into effect. To say that A caused B is to say that B would not have happened without A. But we only have data *with* the policy change. Every statement about cause is really a statement about the way the world would have been without that cause, a *counterfactual* statement. This is one reason why causation is so tricky: it requires reasoning about imaginary worlds that we can never observe directly.

This problem can only really be solved with a time machine. We can go back in time, prevent the new closing time from taking effect, then wait to collect equivalent data in this divergent universe. Lacking a time machine, we'll once again use a model, a way of describing the alternate histories we can't ever observe directly.

If we had two identical copies of New South Wales, we could just change the policy in one city and not the other, and compare the results. This is the logic behind the controlled experiment where you give a new drug to the treatment group and not to the control group. Journalists don't normally get to design experiments, and anyway there are never two identical cities to experiment on. But we could make comparisons with similar cities or neighborhoods.

Just this sort of comparison casts great doubt on an attempt to reduce gun violence in Richmond, Virginia, in the late 1990s. Project Exile aimed to reduce the number of murders by increasing the punishment for illegal gun possession (such as when a previously convicted felon is found to be carrying a gun). The minimum sentence was effectively increased from five to 10 years by shifting all such cases from state to federal courts.

At first glance, it worked.



*Gun homicides per 100,000 residents in Richmond, Virginia, before and after Project Exile. Adapted from Raphael and Ludwig, 2003.<sup>35</sup>*

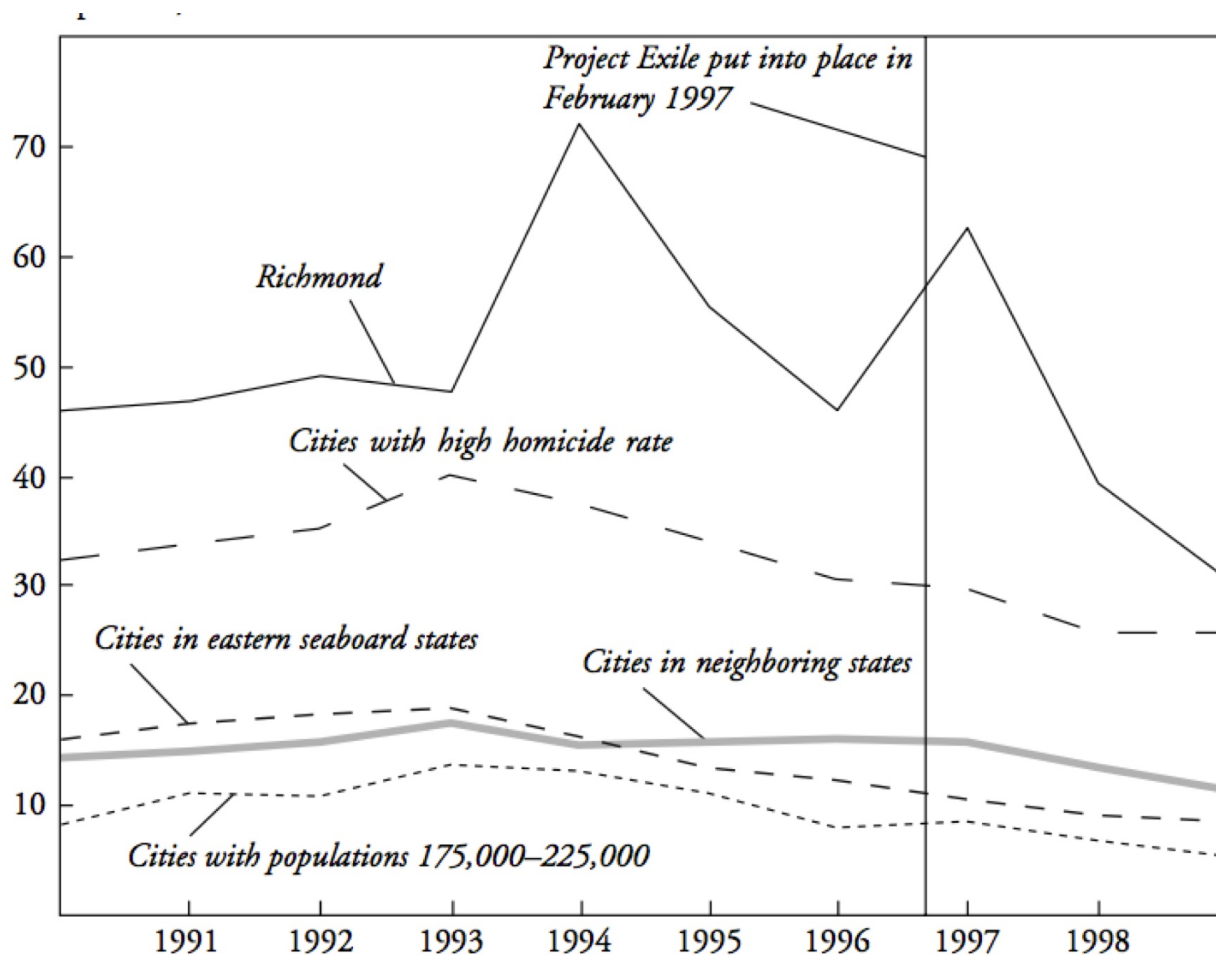
Gun-related homicides—by far the majority of homicides—decreased after Project Exile went into effect. The policy was widely lauded as a success by the National Rifle Association, *The New York Times*, and President George W. Bush.

But the evidence for harsher sentences in Richmond is not nearly as strong as it is for earlier closing times in New South Wales. First, the data is very scarce. There are only three data points after the program was established, for 1997, 1998, and 1999. Further, the number of gun homicides actually increased dramatically for 1997, even though gun possession offenders were tried in federal courts beginning in February 1997. However, 1998 and 1999 do show solid declines, ending lower than anything in the previous decade.

Let's table for a moment the question of chance; with only three data points, luck becomes a real concern. Suppose we believe the decline is real and permanent, and not just fluke due to natural variation. We still have the problem of attributing cause to Project Exile and not something else. Really what we need is another identical Richmond to show us the alternate history where Project Exile never happened.

We don't have another Richmond, but there are many other cities. If those cities are similar enough in the right ways, they might approximate the lost history where Richmond never had a Project Exile. Here's the homicide rate data from other cities which are similar in various ways, but none of which implemented such a program.

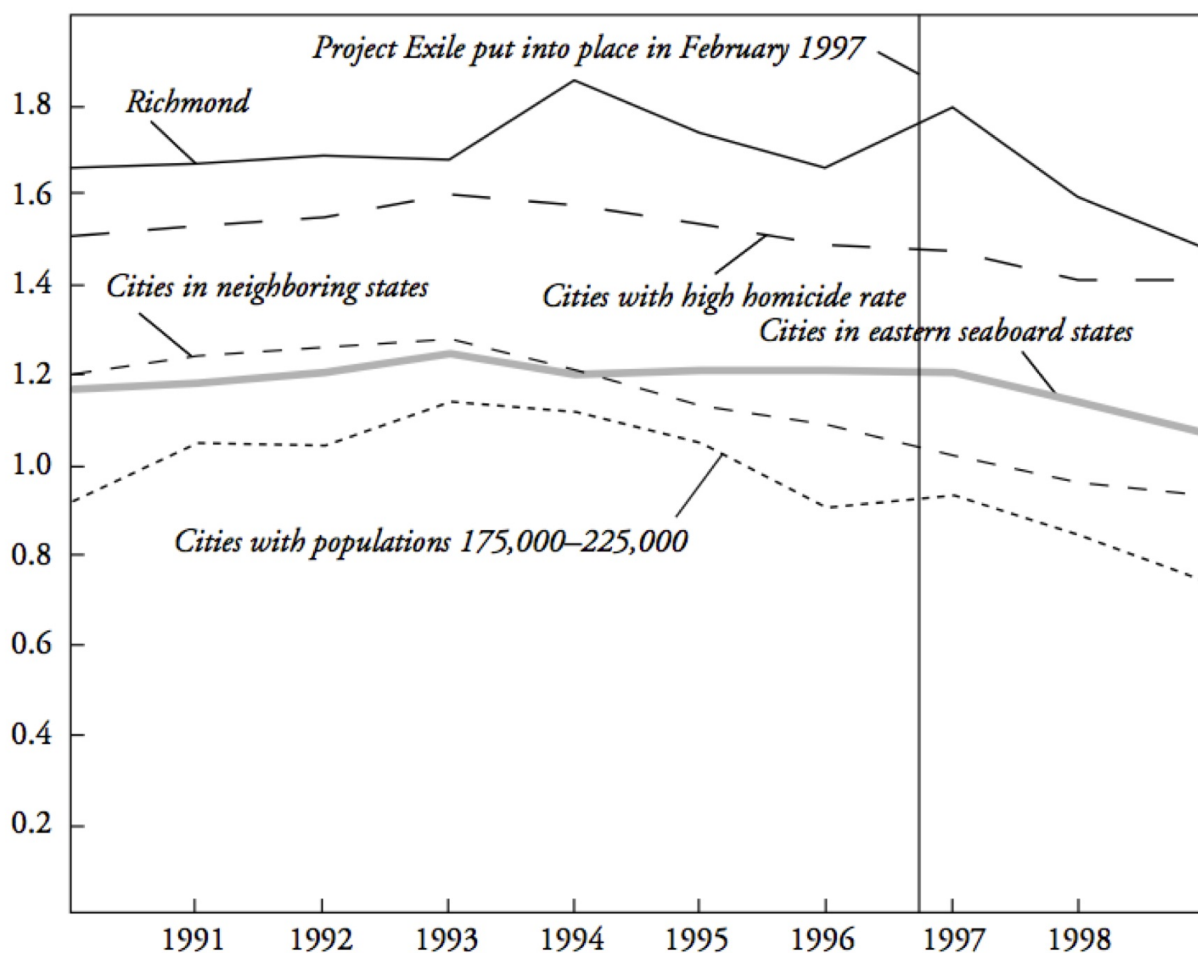




Gun homicides per 100,000 residents in Richmond, Virginia, before and after Project Exile, compared to other cities. From Raphael and Ludwig, 2003.<sup>36</sup>

Virtually every city in the United States experienced a decline in gun violence in the late 1990s. In fact violent crime of all types decreased all through the country during the 1990s. No one really knows why, though there are many theories.<sup>37</sup> Evidently, you didn't need to change sentencing guidelines for illegal gun possession to see a drop in gun crime in the late 1990s.

Maybe you can still say that Richmond had a larger decline. But Richmond also had more crime to begin with, and a big spike in 1997. Proportionally, as a percentage change, Richmond's decrease was well in line with other cities. You can see this if you plot the data on a logarithmic scale.



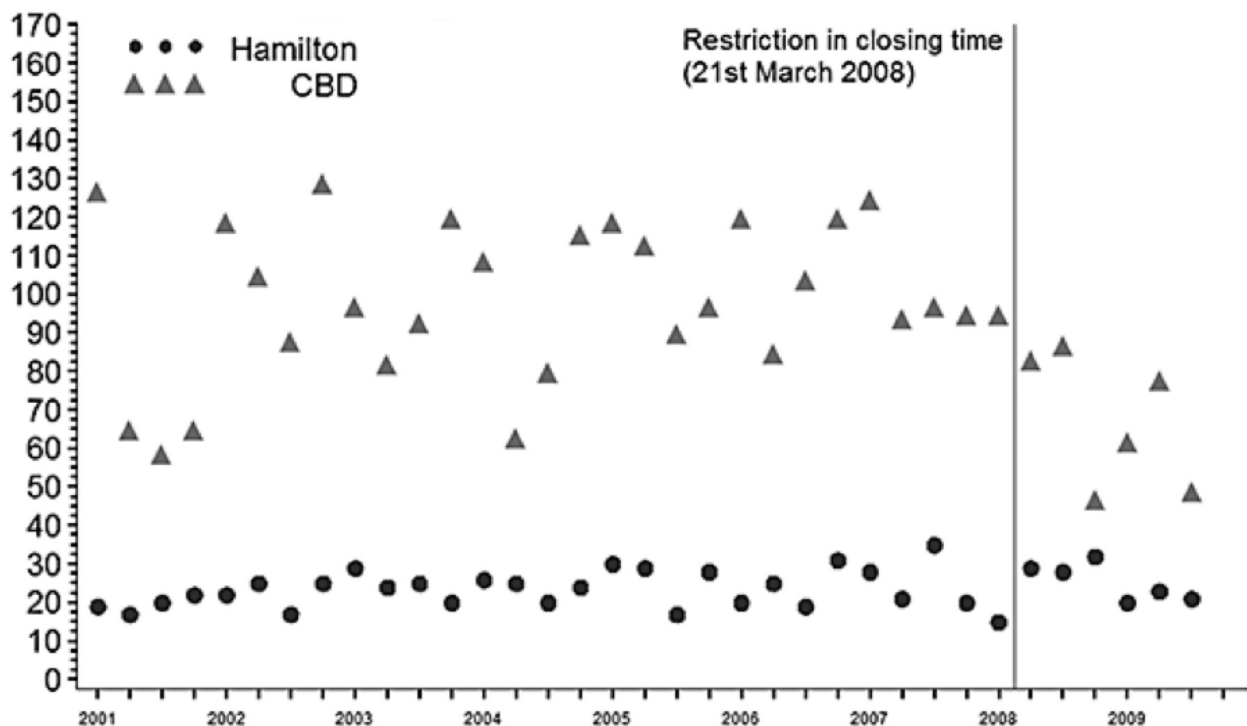
Gun homicides per 100,000 residents in Richmond, Virginia, and other cities, on a logarithmic scale. From Raphael and Ludwig, 2003.<sup>38</sup>

Each vertical step on a logarithmic scale corresponds to an increase by a constant multiplier, which means we are comparing percentage change instead of absolute numbers. When we compare this way, Richmond doesn't look particularly better than other types of cities. Most cities experienced a drop in gun violence of about the same percentage as Richmond, which appears on this chart as a decrease of about the same slope. This is evidence that doing nothing would have been just as effective.

Here you can have an argument about whether percentage change or absolute numbers are the right way to compare a drop in crime between cities. You can also try to construct more elaborate analyses showing that while murders in Richmond would have dropped anyway, Project Exile made them drop more. We're far from the last word, but we're also past a simple argument that Project Exile caused the observed fall.

And, of course, you can jump out of this framing entirely and ask if increased punishment is really the way that we, as a society, want to deal with a type of crime that primarily involves and affects already disadvantaged groups. As always, the data is never the full story.

Back to New South Wales, does the closing-time policy change suffer from the same sort of “would have happened anyway” problem? Again, the theoretically perfect test would require an identical copy of the city. But we do have data from the adjacent neighborhood of Hamilton, which did not see a restriction on closing times.



Number of assaults per quarter in the central business district (CBD) of New South Wales, where closing time was restricted to 3 a.m., and the neighboring region of Hamilton where it was not. From Kypri, Jones, McElduff and Barker, 2010.<sup>39</sup>

And sure enough, there was no apparent reduction in assaults in Hamilton. The main weakness of this sort of comparison is that Hamilton is not perfectly matched with the area where the closing time was changed. It has fewer bars and a far lower rate of assaults to begin with. Still, this comparative data provides a minimal sanity check. We need to exclude the possibility that something *else* happened around the same time that lowered assault rates generally. That’s what seems to have happened with homicides in American cities in the late 1990s. The other reason for looking at the data for the adjacent district is to make sure that crime was actually reduced, not just displaced to nearby areas.

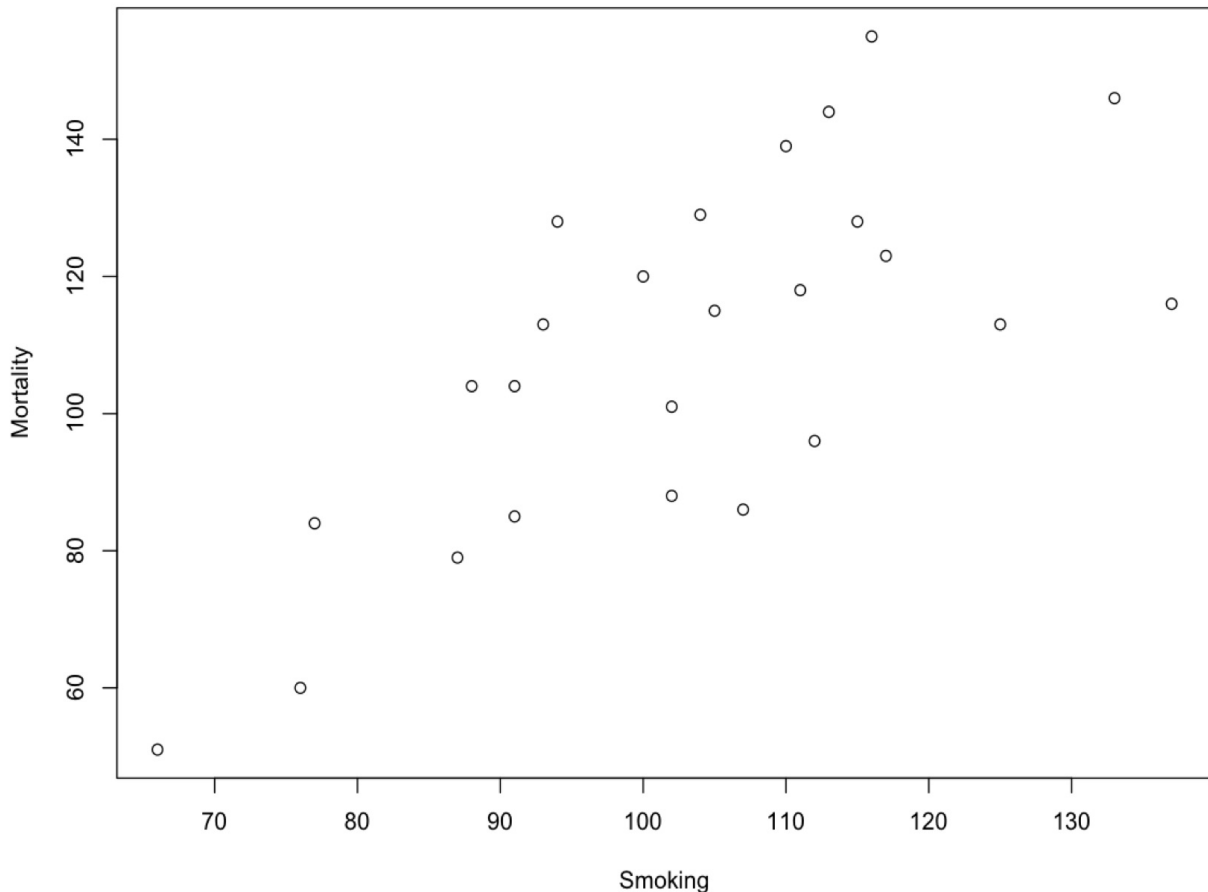
Any claim of cause is implicitly a claim about data from a world we don’t ever get to see: a world where the cause never happened. It’s worth thinking about how to approximate this world through comparisons or modeling. Just looking for increases or decreases is not enough. As the Project Exile researchers put it:

One larger lesson from our analysis of Richmond’s Project Exile is the apparent tendency of the public to judge any criminal justice intervention implemented during a period of increasing crime as a failure, while judging those efforts launched during the peak or downside of a crime cycle as a success.<sup>40</sup>

And that's just not right. The correct comparison is not "up or down," but "what would have happened otherwise?" This applies just as well to the question of whether chicken soup cures colds as it does to the question of whether harsher sentences deter crime.

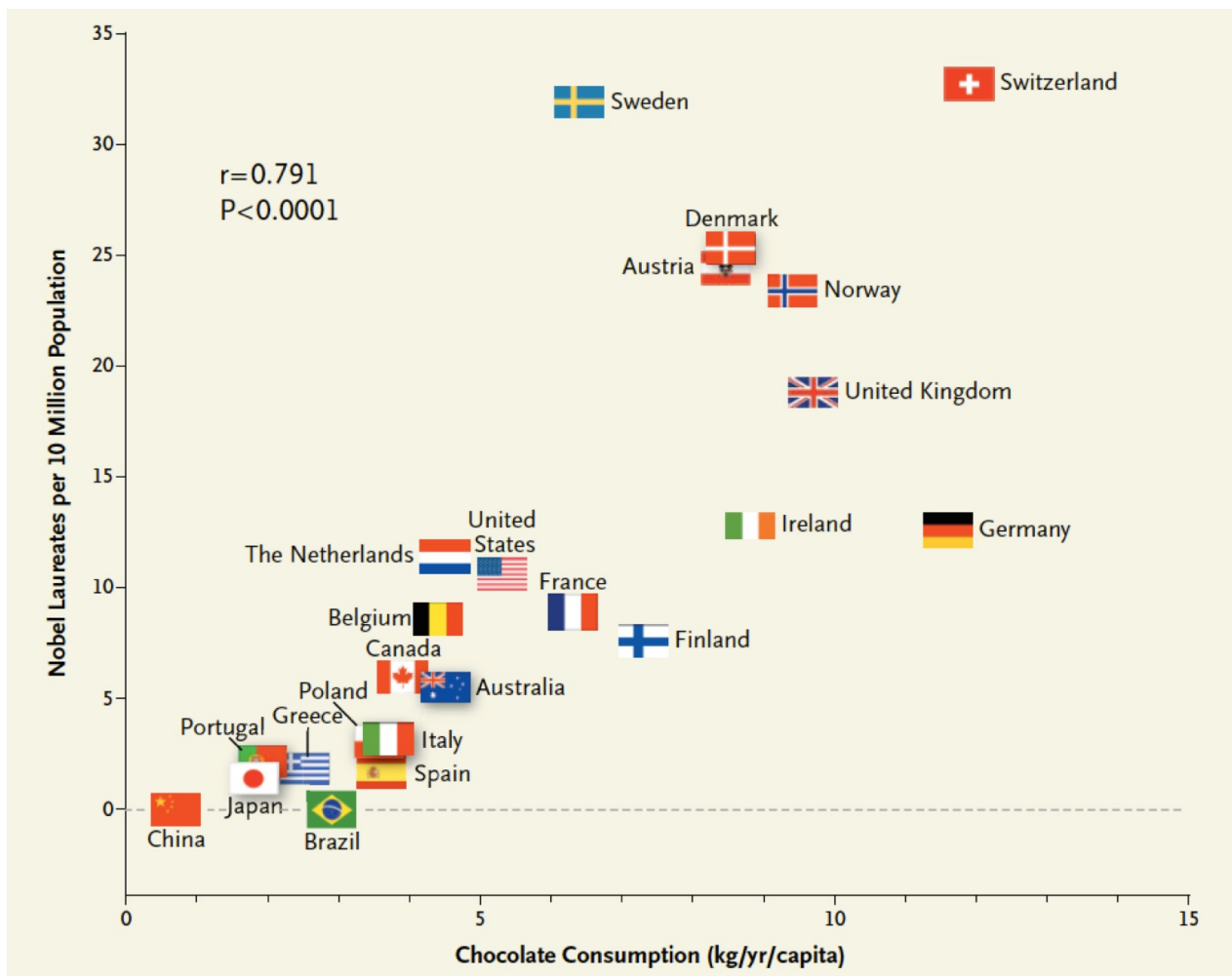
## Causal Models

Cause cannot usually be read directly from the data, no matter how much we might wish this were the case. Consider this graph of mortality versus smoking rate across different occupations:



*Normalized mortality rate versus smoking rate for different professions in the United Kingdom, 1970–1972.*<sup>41</sup>

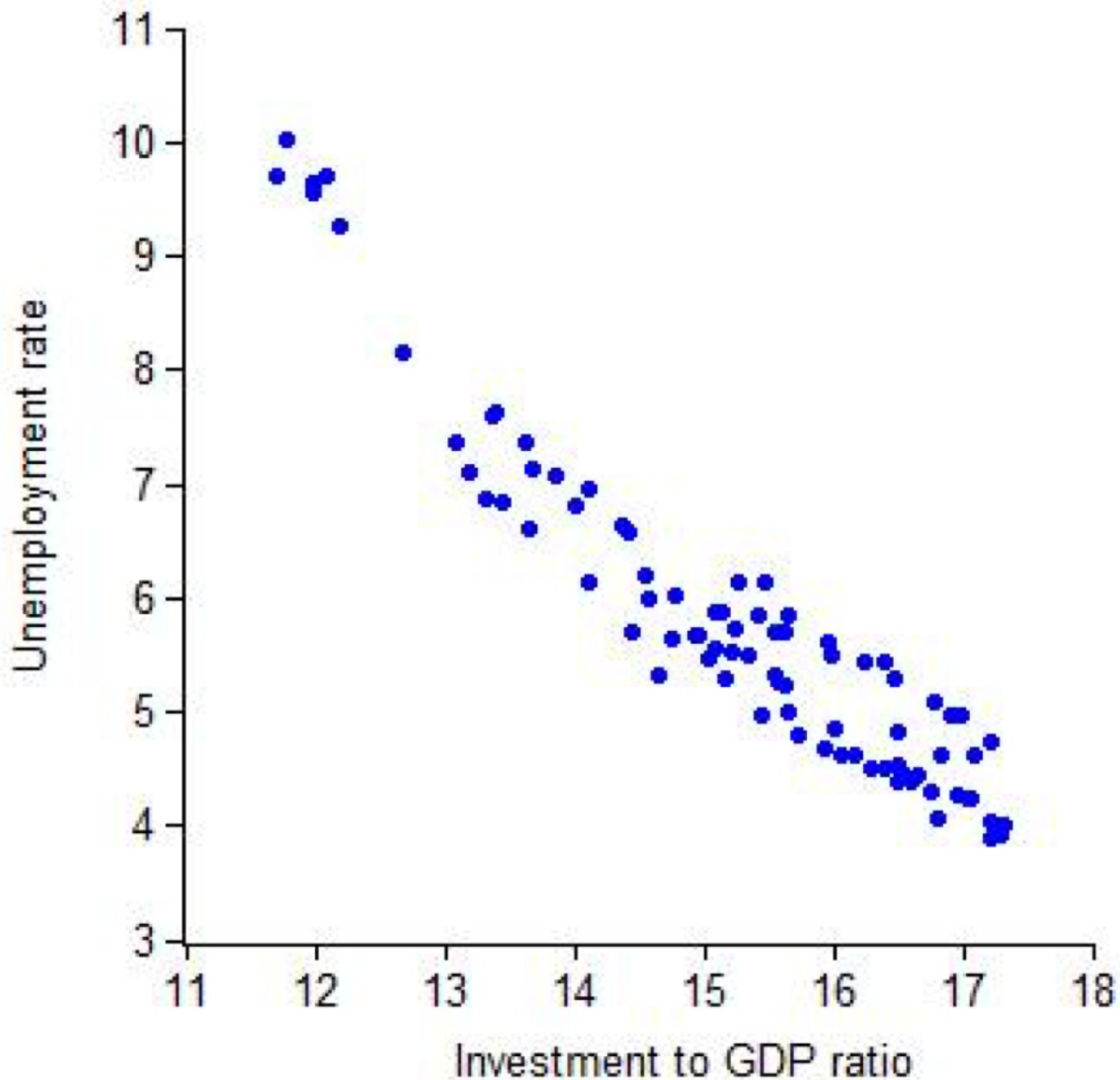
There is a clear association between smoking and mortality—a correlation. It seems natural to say that this is evidence that smoking contributes to an early death. But how about this chart:



Correlation between countries' annual per capita chocolate consumption and the number of Nobel Prize winners. From Messerli.<sup>42</sup>

If the previous chart shows that smoking causes premature death, then this chart shows that eating chocolate makes you more likely to win a Nobel Prize. No? But then why do we believe the first correlation is causal, while this one isn't? There must be some other factor here; our reasoning must be including something other than just the data.

Here's a more ambiguous case:



*U.S. quarterly unemployment rate versus investment to GDP ratio from 1990 to 2010, plotted by John Taylor.*<sup>43XX</sup>

How would you describe this graph? Maybe: When investment goes up, unemployment goes down. But saying it that way makes it sound like increasing investment would cause unemployment to drop, and that's not necessarily true. We might as well say that when unemployment goes down, investment goes up, implying a cause in the other direction. Perhaps we could say: Investment and unemployment move together, in opposite directions. That's all we actually know from this data, yet it feels unnatural to write about an association between two variables while saying nothing about the causal relationship between them. We are wired to see causes.

The difference in our intuitions about these three charts has to do with whether or not we know a story that explains how the cause relates to the effect. You can probably imagine how investment would lead to employment, or perhaps how employment would lead to

investment. You've also probably heard that smoking causes cancer. But there's no obvious story that links eating chocolate and winning a Nobel Prize.

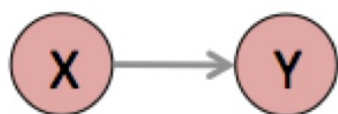
We are dealing with a *correlation* here, a pattern in two variables such that when one changes the other changes as well. There are various mathematical definitions of a correlation, but for our purposes the most straightforward conception is fine. Scatterplots are a popular way to compare two variables, but anything which shows two variables can reveal a correlation. One of those variables might implicitly be the time of an event, as in our crime examples where we were looking at the correlation between a change in policy and the number of assaults or murders. Here's another type of correlation, from an analysis of men writing a first message to women on the dating site OKCupid:<sup>44</sup>



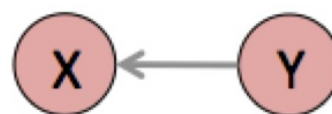
This data seems to show that including the word “awesome” in a first message will *cause* an above average reply rate, while including the word “sexy” will *cause* a much lower chance of a response. But that's not what the data actually says. That's just a story that leaps to mind. It's easy to imagine why women would ignore a creepy first message from a stranger who called them “sexy.”



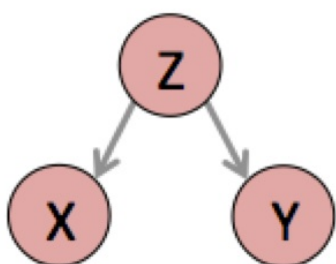
As usual, our stories about the data may or may not reflect reality, and the principle method of testing our stories is trying to imagine how else the data might have come to be. Fortunately, there are not that many ways two variables can become correlated.



**X causes Y**



**Y causes X**



**Z causes X and Y**



**random chance!**

These little graphs are causal models. Like all statistical models, they are not reality but a way of talking and thinking about reality. Each circle is a variable, something that is or could be quantified. Each little arrow means “causes.” What exactly a “cause” is has been debated since Aristotle, but in this framework it is defined in terms of possible interventions: X causes Y means that there is some specific thing you could do in the world to force the variable X to take a specific value, and if you did that the outcome of Y would change in a probabilistic sense.

These causes are not definite. To say that smoking causes cancer means that if you could force someone to smoke, they would be more *likely* to get cancer. Not that they *will* get cancer, but that it increases the probability. The arrows in these diagrams are fuzzy, probabilistic cause. Instead of “causes,” think “changes the distribution of.”

This level of abstraction lets us talk about cause in a very general way. Every correlation of any two variables is the result of one of these causal patterns, or more likely a combination of them. Usually, the data alone cannot tell you which pattern produced your correlation.<sup>xxi</sup> For example, X causes Y and Y causes X appear the same in the data. We have to use other information to figure out the correct causal structure.

There could be no causal relationship at all, just random coincidence between X and Y. As we've seen, coincidence can be quantified by estimating the probability that chance generated your data. In the OKCupid case we could ask: How often does a randomly chosen word have an above- or below-average response rate as large as these words? If we plot the response rates of lots of words, we may find that these particular words are not special at all; this chart could just show some particularly entertaining words that have quite ordinary fluctuations in response rate. If you can cherry-pick the evidence, you can prove whatever you want.

It can also be that Y causes X, but not in this case. The reply cannot cause the initial message because causes have to come before their effects. In other cases the causality could flow in the other direction, or the variables could affect each other in a feedback loop. High unemployment might be both a cause and effect of low investment. If cities with more guns are associated with higher crime, it could be that access to weapons causes crime, or it could be that living in a dangerous place makes people want to buy a gun. Or the association could have happened purely by chance.



**if you have a gun, you're going to use it**



**if it's a dangerous neighborhood, you'll buy a gun**



**the correlation is due to chance**

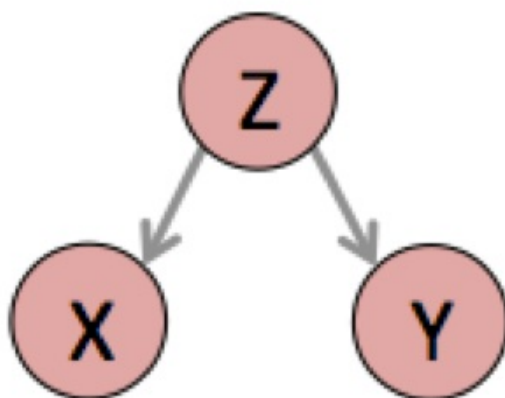
In reality, it's probably some combination of all of these effects. The data you have is the result of people using the guns they have *and* people buying guns because of the high crime rate *and* a whole range of chance factors.

It could also be the case that some other factor Z causes both X and Y. For example, there could be something that causes a man to write about a woman's appearance *and* causes a woman to reply less often. This is the possibility most often neglected in casual data analyses, but there could be any number of factors that would influence both language use and response rate.

Like attractiveness. Perhaps attractive women get a lot more messages than average—too many to want to reply to all of them—so their overall response rate is lower. If we believe that “attractiveness” is a real and coherent notion that could be usefully measured in some way—perhaps by asking many people to rate a photograph—then it is reasonable to talk about it as a variable. This leaves us with two plausible hypotheses.



**telling a woman she's beautiful  
makes her respond less**



**if a woman is beautiful,  
1) she'll respond less  
2) people will tell her that**

There is no way to tell these two hypotheses apart from the data above, because both would produce the same correlations.

The third variable in this three-way structure is called a *confounder*, and confounding variables appear frequently in real world analyses. The key is to look for another variable that causes *both* of the variables you see as related. For example, overall economic growth could both reduce unemployment and increase investment. A rich country might both import a lot of chocolate—a luxury good—and fund advanced research. The reduction in crime rates after the bar's closing time changed could be because the police began patrolling to enforce the earlier closing time.

But then again, a stressful profession could both make you smoke and reduce your lifespan. The tobacco industry has attacked the association between smoking and disease for decades on precisely this basis of possible confounding variables (and many other arguments<sup>45</sup>). In the mid 1960s, one statistician received tobacco industry funding “to seek to reduce the correlation of smoking and diseases by introduction of additional variables.”<sup>46</sup> As repugnant as this might be, we have to take seriously the logical possibility of a spurious correlation. Ultimately, the proof of smoking's harm also relies on other types of non-correlational evidence such as animal experiments. We can tell a story about smoke causing cancer that we can confirm in the lab.

Confounding variables are common in practice. Coffee might cause cancer, but then again maybe a certain type of person both smokes and drinks coffee.<sup>47</sup> Poor sleep might cause poor grades in school, or poverty might cause both.<sup>48</sup> The confounding circumstance may not be measured in the data you have and may not even be something that can be measured directly. You can only find a confounder by thinking about the broader context of the data.

Once you have found a confounding variable, it may be possible to subtract off its effect, a process that is called *controlling* for a variable. For example, you could investigate the relationship between smoking and cancer while controlling for the stress of different professions. This only works if your causal model is otherwise accurate. Again, it's a way to ask about a counterfactual: What would be the relationship between employment and investment *if* growth didn't drive both of them? Or how much would women make *if* they worked the same number of hours as men? Reasoning about imaginary worlds is always tricky.

I've used pictures informally to talk about causal structures, but they're actually part of a well-founded mathematical theory of cause developed in late twentieth century by Judea Pearl and others.<sup>49</sup> These pictures are called *graphical models*, not because they are

graphics but because they are graphs in the mathematical sense of nodes and edges. You can use them to describe much more complex causal structures with more variables, like this model from one of my favorite statistics books:

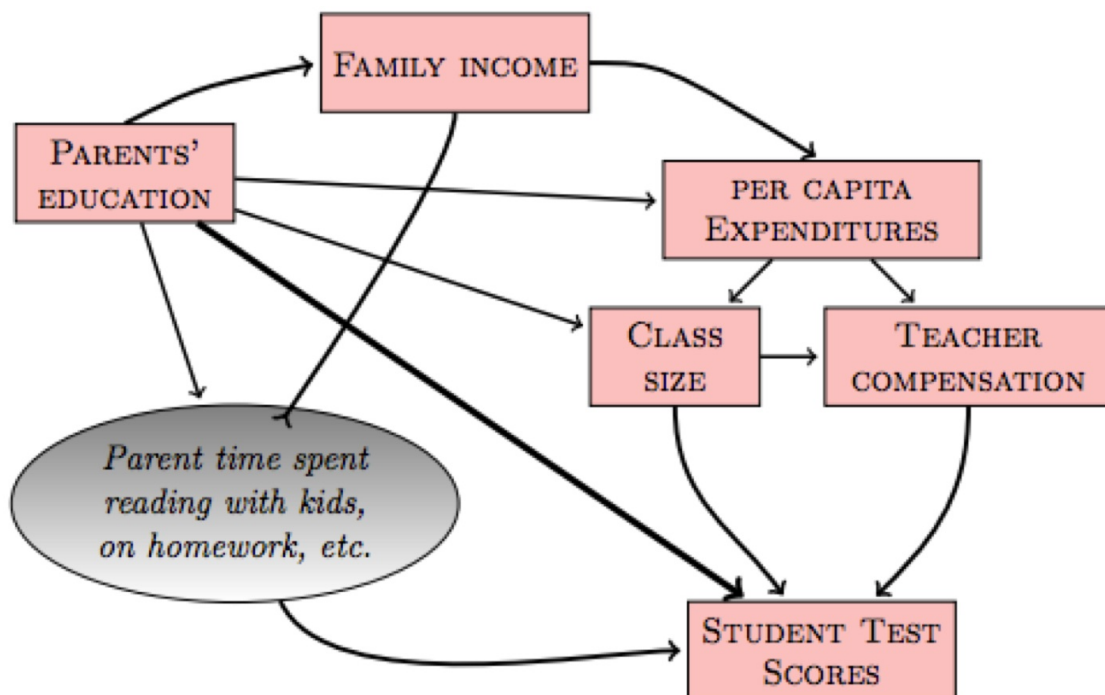


Figure 17.9: A hypothetical causal network relating student test scores to various educational policies and social conditions.

From Kaplan.<sup>50</sup>

In this invented network we have data for the pink variables but not the gray variable. In general there will be many intervening factors you can't measure, as well as unknown causes that you may never have thought of. You just don't know the correct causal structure of the world, but at least you can draw little pictures of the possibilities you can imagine.

The best way to figure out causation is to do an experiment. After all, causation is defined in terms of interventions, and an experiment is all about intervening. In the online dating case, we could take many men and randomly tell each one to include or exclude certain words in their first message to a woman, then tally the response rate for each word. This is different from the data we already have in a crucial way. In this experiment the men do not decide which words to use (we have intervened!). They cannot base their decision on the woman's appearance, or for that matter anything about themselves or the woman to whom they are writing. This removes the effect of many potential confounding variables in one shot.

This type of experiment is a generalization of the idea of comparing cases. We repeat a particular scenario many times with and without the hypothetical cause and see if the effect appears more often when the cause is present. John Stuart Mill wrote about this "method of

difference” in his 1843 *A System of Logic*:

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.<sup>51</sup>

Mill understood that it would not always be possible to distinguish “X causes Y” from “Y causes X” from data alone (“is the effect, or cause”). Experiments are one way out, because we set the value of X and watch what happens to Y. The hitch is that we don’t know what would have happened to Y if we didn’t set X. How many non-smokers would have developed lung cancer anyway? This is why modern experiments use a *control group* for comparison. To ensure that the two groups are otherwise identical (“every circumstance save one in common”), we can randomly assign people between them. This basic design was formalized at the end of the nineteenth century and is known as a *randomized controlled experiment*.

But again, journalists don’t normally get to do experiments. Sometimes we can evaluate other people’s experiments, but usually we are reduced to dealing with *observational* data. This makes cause an especially tricky subject. Causal models—our little arrow diagrams—are a way of expressing the possible causal relationships between variables. This can clarify our thinking and hopefully lead to ideas about how to test our stories against reality.



## Truth by Elimination

In 2011 the Associated Press revealed that the New York Police Department had been closely monitoring 53 New York City mosques with methods including informants and video surveillance.<sup>52</sup> In 2012, the NYPD released a massive database of hundreds of thousands of stop-and-frisk incidents, where cops stopped people on the street, without cause, to check for weapons and drugs. A journalist analyzed this data and found that there was a 15 percent above average number of stop-and-frisks within 100 meters of certain New York City mosques.<sup>xxii</sup>



*A small portion of the NYPD's stop-and-frisk data.*

This might mean that the NYPD is deliberately targeting Muslims on the street. But there are many other ways this data could have come to be. Let's list some possibilities:

- Police are deliberately stopping Muslims near mosques.
- It's sheer chance.



- Mosques could be in more heavily populated areas.
- Patrol times might coincide with prayer times, for whatever reason.
- There might be more police assigned to the area due to higher crime rates.
- The data might be in error.
- You could misunderstand how the data is collected.

This is the central problem of data analysis: The data alone cannot tell us that a story is true, because there could be many other stories that produce the same data. In principle all scientific analysis is a two-step process: Invent a number of hypotheses, then pick the one which is best supported by evidence. In journalism work, a narrative extracted from the data—“the story”—is morally equivalent to a hypothesis.

Actually, neither scientists nor journalists really work like this. Many people have pointed out that the interplay between inventing and testing ideas is much more complex than this little sketch.<sup>53</sup> In real work you go back and forth, refining ideas, gathering more information, finally getting your interview with a crucial source, testing theories, catching up on other people’s work, stumbling into flashes of creativity, drinking a lot of coffee, arguing with critics, going back to the drawing board, changing your mind, grinding forward. We should not consider this idea of creating and then testing hypotheses to be a literal description of our truth-finding process. Instead it describes a type of argument. It captures the core logic of why we should believe something is true, not necessarily the steps that actually led us to believe it.

Coming up with reasonable stories/hypotheses is a creative process that has to draw on specific background knowledge. Peirce called this hypothesis-generation process abduction and noticed that it followed certain rules: Your stories must explain the data, and they must not contradict known facts. Other than that, the possibilities are wide open. But there are a number of things that need to be checked in almost any story. Your list of hypotheses should include definitional problems, quantification troubles, errors in the data, random chance, and as many confounding variables as you can think of. The basic rule is this: you have to imagine it before you can prove that it’s true.

Is NYPD targeting of Muslims producing our data? The truth may be any of the possibilities above, some combination, or something that’s not even on the list.

If you have well-quantified variables and good models, there are statistical solutions to the problem of choosing between competing hypotheses. Much of the statistical work of the last hundred years has been devoted to just this sort of hypothesis testing, as we saw in the section on inference. These are powerful tools, but most problems in journalism do not have neatly quantified evidence. I don’t know how to express all of the

above stop-and-frisk hypotheses in the same symbolic language, nor how to make reasonable probability estimates for each possibility. What's the chance you've misunderstood the data format? In practice the solution is to double-check the format, rather than trying to compute a probability of error.

There are exceptions, highly structured cases where the full power of statistical hypothesis testing can be applied, such as election predictions. Even then, be wary: Have you included all the different ways the election could be rigged? The world will always find ways to surprise a model.

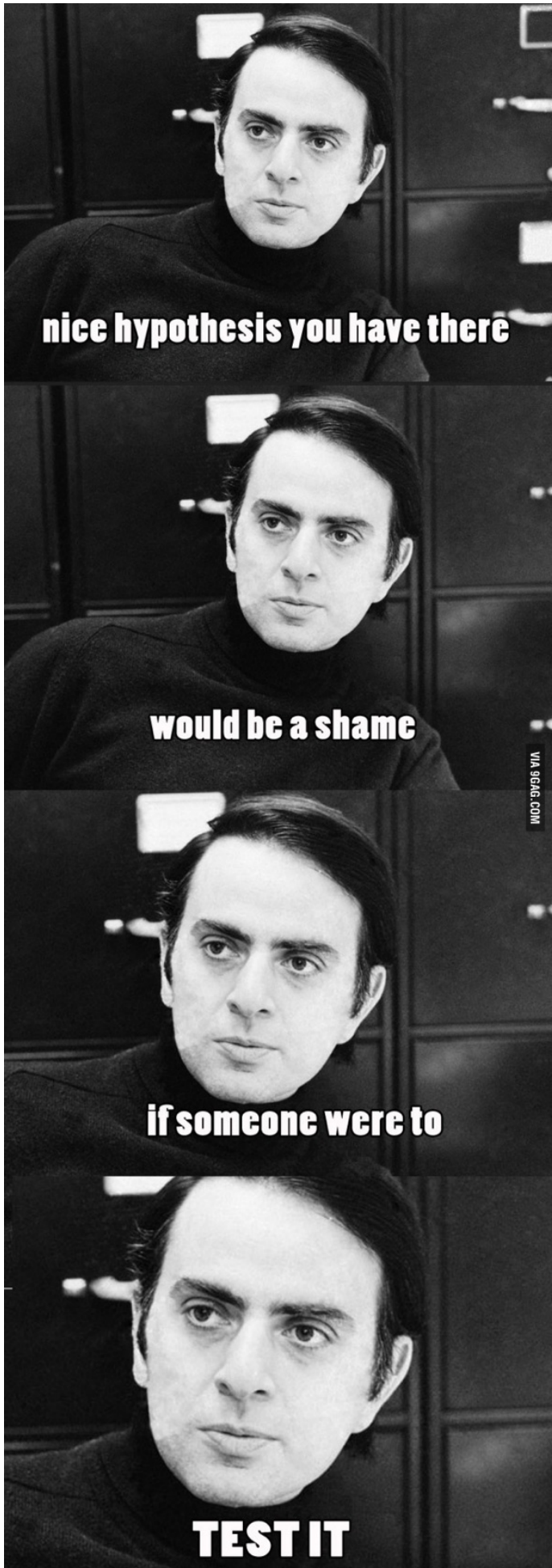
Ultimately there is no language more powerful than human language, and no reasoning more powerful than general human reasoning. That doesn't mean looking at the data and intuiting the answer. There are many methods between intuition and statistics.

Good data analysis is more about ruling out many false interpretations, rather than trying to prove a single interpretation is correct. This may seem disappointing—can there be no certainty?—yet this idea is one of the great innovations in philosophy of science. It was best articulated by Karl Popper in the 1930s. His central idea was that *falsification* is a much more robust practice than *verification*.

There are many reasons why proving a story wrong is a better goal than proving a story right. If you only ever look for evidence that confirms your story, you may only ever find the evidence that confirms your story. Disconfirmation is also more powerful than confirmation in the sense that additional confirming evidence doesn't really make a confirmed story more true, but once a story is contradicted by a single solid fact no amount of further evidence can rescue it. And we know, starting with a series of landmark cognitive psychology experiments in the 1970s, that there are biases in human cognition that lead us to reject, discredit, and selectively forget information that doesn't fit with what we already believe.<sup>54</sup>

It's useful to inquire against your hopes. Your critics certainly will.

Also, falsification is a way of clarifying the practical content of a hypothesis. Is there some way, at least in principle, that your hypothesis could be proved wrong? If a hypothesis says anything about the world, it should be possible to go check if the world really is that way. I don't mean anything cosmic by this. "The police shift change happens during evening prayers" is a perfectly good hypothesis that could be tested by, say, getting a copy of the precinct schedule.



Carl Sagan throws down.<sup>xxiii</sup>

The idea of generating competing hypotheses and then disproving them appears in many forms, in many places. Aristotle wrote about the idea of different possible causes for the same event. Peirce certainly understood the principle in 1868 when he used his signature model to rule out chance as an explanation. Sir Arthur Conan Doyle had Sherlock Holmes talk about finding truth by testing alternatives in 1926, in the quote that opens this chapter. A 1980s CIA textbook on intelligence analysis contains a particularly readable description of a practical method, neatly tied to the theory of cognitive biases.<sup>55</sup>

In short, the method is this: At the beginning of the data analysis work, dream up all sorts of possible interpretations, all sorts of possible stories. The available data will rule some of them out, either obviously so or through statistical testing. The stories which survive that test are the ones you have to choose between. To do that, you will need more information. The remaining set of hypotheses will tell you which information you need to rule each of them out, whether that's another data set or a conversation with a knowledgeable source.

Each of the stop-and-frisk hypotheses suggests a different investigative technique. We can examine the effects of chance statistically, perhaps by counting the number of stops within 100-meter radius circles placed randomly throughout the data, not centered on mosques at all. But pretty much every other hypothesis has to be tested against information that isn't in the stop-and-frisk data. We might want to add other data to the analysis; for example, we could correlate mosque locations with population density. Or we might need to have a conversation with a cop who can explain how police patrols are assigned. The goal here isn't to prove any particular hypotheses but to test each of them by finding evidence against them.

We're looking for information which falsifies one of our hypotheses. Reality may not be so cooperative. The next best thing is information which prefers one hypothesis to another: not falsifying evidence but differential evidence. We might also find that a combination of hypotheses fits best: The NYPD might be intentionally stopping Muslims on the street *and* mosques might be in more densely populated areas. That itself is a new hypothesis.

The method of competing hypotheses need not involve data at all. You can apply the idea of ruling out hypotheses to any type of reporting work, using any combination of data and non-data sources. The concept of *triangulation* in the social sciences captures the idea that a true hypothesis should be supported by many different kinds of evidence, including qualitative evidence and theoretical arguments. That too is a classic idea. Here's Peirce again:

Philosophy ought to imitate the successful sciences in its methods, so far as to proceed only from tangible premises which can be subjected to careful scrutiny, and to trust rather to the multitude and variety of its arguments than to the conclusiveness of any one. Its reasoning should not form a chain which is no stronger than its weakest link, > but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected.<sup>56</sup>

What you see in the data cannot contradict what you see in the street, so you always need to look in the street. The conclusions from your data work should be supported by non-data work, just as you would not want to rely on a single source in any journalism work.

The story you run is the story that survives your best attempts to discredit it.

# Communication

*The mark of a civilized human is the ability to look at a column of numbers, and weep.* - attributed to Bertrand Russell<sup>xxiv</sup>

Quantification produces data and analysis brings meaning to it. But it doesn't count as journalism unless you can communicate what you've learned. This need shapes the story all the way through, including quantification and analysis.

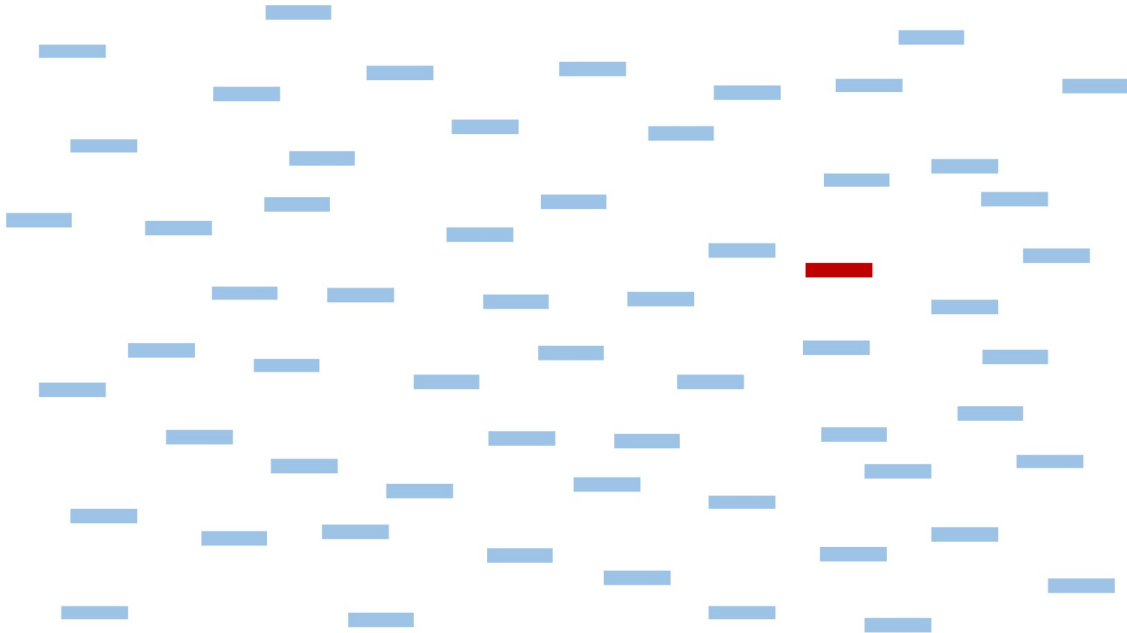
In journalism we usually need to assume that the audience has little familiarity with either the subject of the story or quantitative concepts in general, which makes this particularly difficult. And after reading, the reader<sup>xxv</sup> must eventually do something with the information, or our journalism has no effect. This ties journalism to prediction.

Most people are not used to interpreting data, and it's hard to blame them. Data visualization can be helpful because it transfers some of the cognitive work of understanding data to the enormously powerful human visual system. Still, the foundational concepts of data work are subtle and at times unnatural. The nuances of sampling, probabilities, causality, and so on are foreign to everyday experience. More than that, numbers are not a particularly empathetic medium. For most people even the most screaming statistic is disconnected from everyday experience. Journalists can overcome this using examples, metaphors, or stories to relate the numbers to people. Journalism is a deeply human task, no matter the methods.

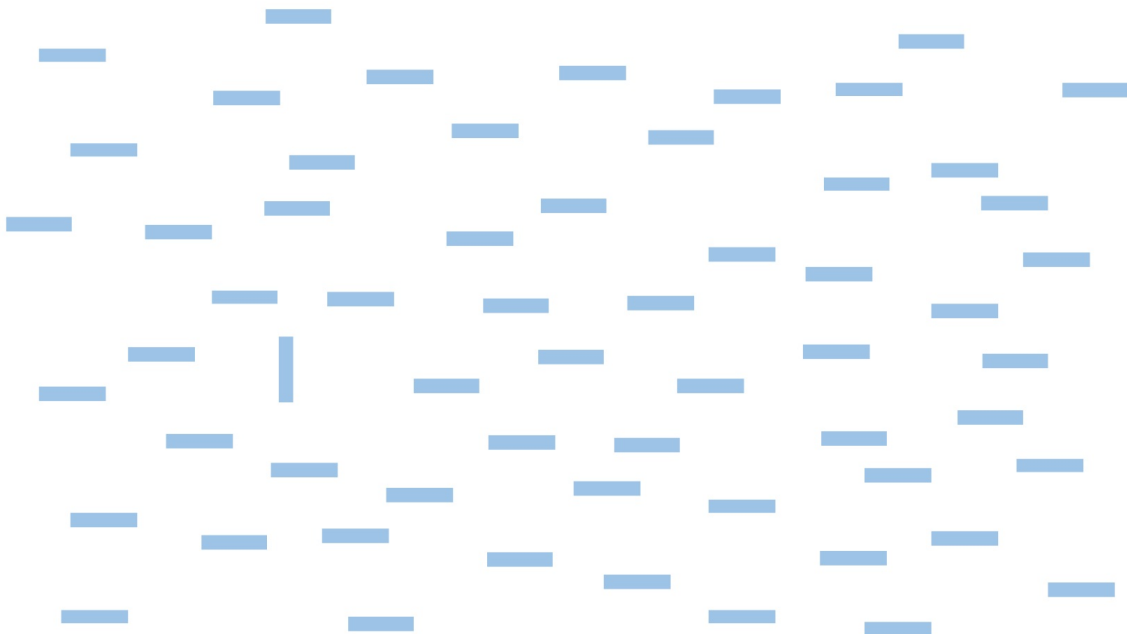
Ultimately, a journalist is responsible for the ideas that end up in their reader's head. There are two parts to this: ensuring that the data and the story clearly and accurately represents the reality, and ensuring that this accurate representation is what the reader actually comes away with.

# Perception

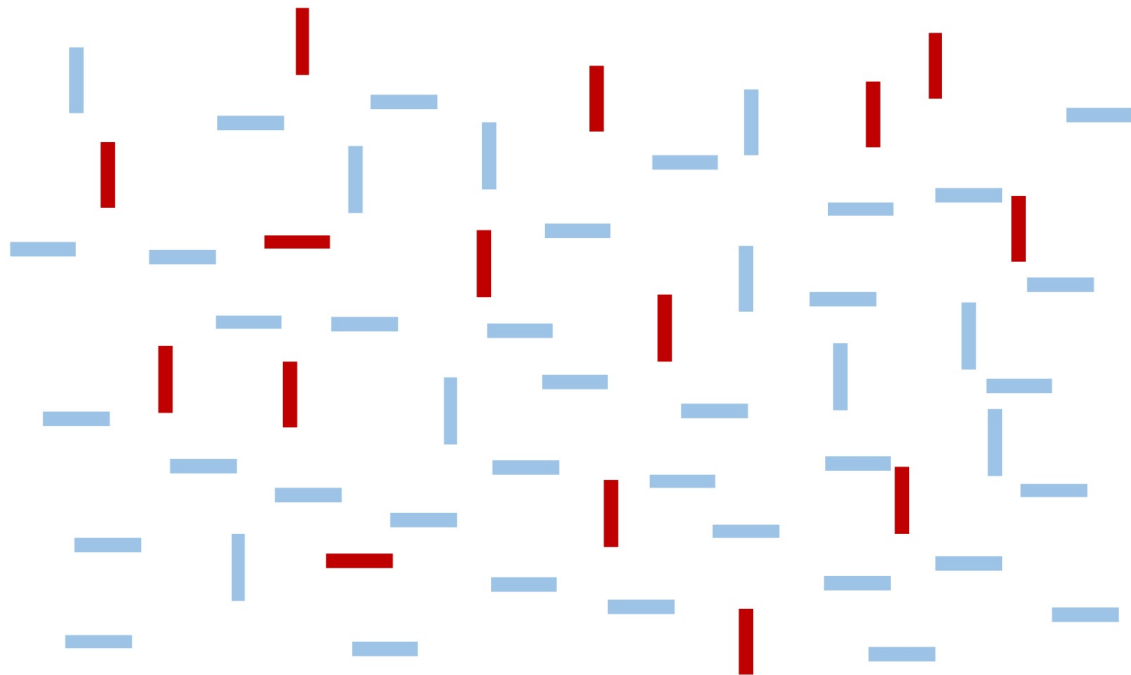
Quick, which of these shapes is different?



Well that was easy. How about now?



Now try this one. Which shape is different from all others here?

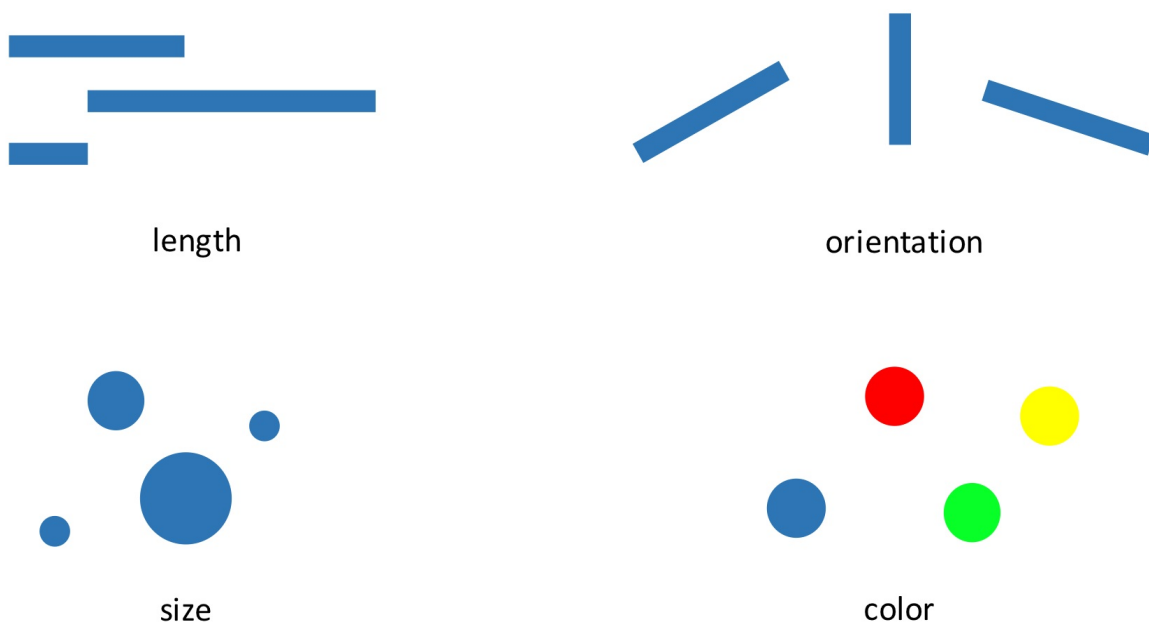


The first two were easy, but that one was slightly harder, right? These examples illustrate a visual ability called the pop-out effect, which lets you find something in a sea of similar objects without having to think about it. The object that is different just “pops out” at you. Except that sometimes it works better than others. You probably took a few seconds longer to find the single vertical light bar in the last image.

Pop-out sometimes works and sometimes doesn’t because you have “hardware” in your visual system that can perform complex processing tasks below the level of your consciousness. Under the right circumstances, color, orientation, shape, texture, motion, depth, flicker, and many other visual attributes can cause pop-out. But if the problem gets too complex for your highly specialized visual hardware, you have no choice but to do a “visual search” by scanning each object, like a Where’s Waldo book.

Your visual system can do all sorts of other neat tricks, like comparisons.





You don't have to think to know which object is largest, or tilted down the most, or whether the circles are different colors. This is the basis of all data visualization: We are relying on very rapid, unconscious abilities of the human visual system to communicate data quickly. With a well-designed visualization, you don't need to think about it to see a trend or a cluster.

Data visualization researchers have identified many important features of the human eyes and brain.<sup>57</sup> There are different visual "channels" we might use to encode data, such as position, size, color, orientation, shape, texture, motion, depth, and a dozen more, and from experiments we know the effectiveness of these channels for different types of representation. For example, we know that position is the fastest and most accurate visual channel for comparing quantities, while color works great for categorical data but poorly for continuous variables. We've measured how perceived contrast changes depending on context, and explored how noise and clutter can slow down visual tasks. And we've teased out how pictures save on short-term memory. With a picture in front of you, you don't need to store the relationships between elements in your working memory, because you can just look and see. This frees up your thinking for more sophisticated thoughts about the content.

Our visual processing system is so fast and sophisticated that maybe we shouldn't think about it as cognition at all. Instead, it's perception. It feels like you "just see" the important features of the visualization. But of course we don't "just see." Experimenters have mapped out exactly what we do and don't see, and you can train your eye over time, too—like when you learned to recognize letters and then words.

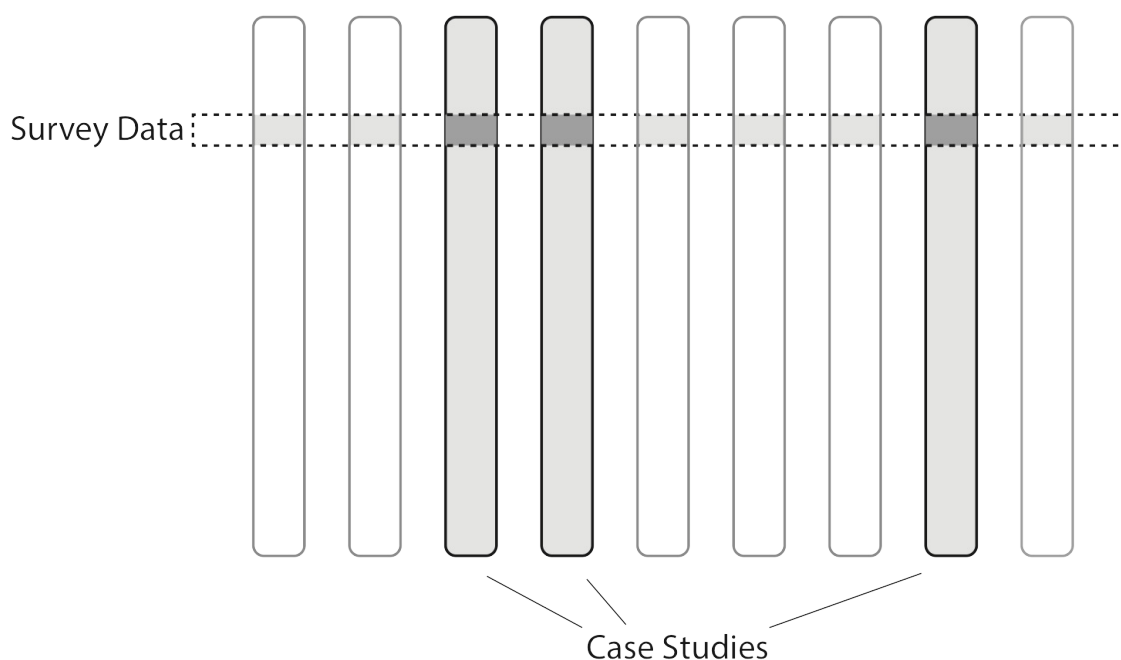
Considering our visual abilities leads to important design choices. Our unconscious ability to compare lengths is why you should generally start the Y axis at zero. Otherwise, the relative lengths won't correspond to the relative values, and we'll perceive incorrect relationships. Ignoring visual perception when creating data visualizations is like ignoring the consensus meanings of words when writing.

But it's not just vision we need to understand. We can't possibly study the communication of data without studying the human perception of quantities. How our story is perceived depends on everything from vision to cognition to what the audience already believes.

## Representation

Most of what we know comes through some form of media, some form of secondhand representation. A great deal has been said on who and what gets represented in journalism, and how certain people and ideas are presented. Adding data does not change the basic nature of these issues, but data is a different kind of information that lends itself to different kinds of communication.

I tend to think of information as coming in two different flavors: examples and statistics. The story of someone looking for a job is an example, while the unemployment rate is a statistic. People also talk about anecdotes versus data, or case studies versus surveys, or narratives versus numbers, or maybe qualitative and quantitative. Not all of these pairs are talking about quite the same thing, but they all capture some kind of difference. I don't think these modes of information are in opposition, or even that the boundary is really all that clear. (What would you call the ethnographies of a randomly sampled set of people?) But I do see two very general patterns in the way information can be collected.



You can collect a small amount of specific information from many people and summarize it with statistics. Or you can collect rich, open-ended information from just a few people and present each as an in-depth example. In this sense statistics and examples are

complementary forms, and both can be used to represent a broader group of people. That is, both can be used to infer information we did not collect—additional details about the lives of more people. All representation is generalization.

Consider unemployment again. A survey asks a few questions of many people, so that we can count how many people are unemployed. We can also find patterns of connection between employment status and location, education, age, and so on. To see these patterns truly, without bias, we must either count every single person or take a random sample. That is, a random sample is a *representative* sample. But we also need to understand the lives of individual people, or we cannot ever understand how these societal forces play out in practice. Maybe we know that people of a certain race have higher unemployment, but how does this actually happen? What goes on in such a person's life when they are looking for a job? What did they hear in their last interview? The unemployment rate cannot answer these sorts of questions, but the stories of individual people can.

In the best case, a story combines numbers and narratives. The data represents many people in a narrow but meaningful way, while stories relate the deep experiences of only a few, and these different types of information together describe a unified reality. But this is only what's on the page.

## Examples Trump Statistics

Taking responsibility for the impression that the reader comes away with requires an understanding of how people integrate different types of information. And generally, examples are much more persuasive than statistics—even when they shouldn't be.

The United States has seen a two-decade-long decline in violent crime rates. This holds across every type of violent crime and in every place.

### *U.S. Violent Crime Rate, U.S. Justice Department Statistics, 1973-2010*

Number of victims per 1,000 population aged 12 or older



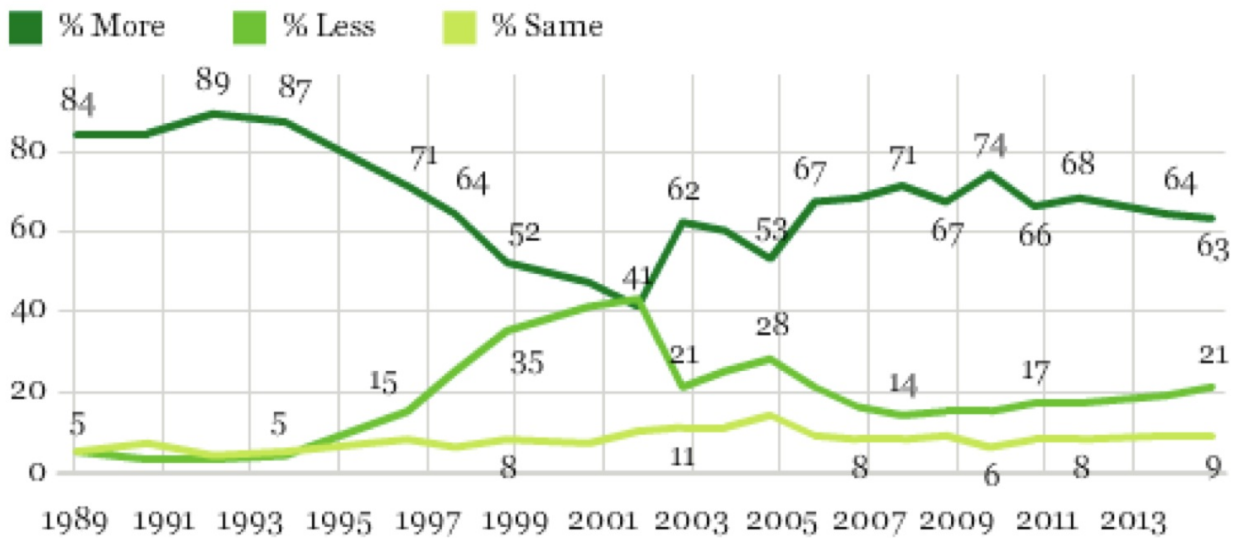
Source: Bureau of Justice Statistics

### GALLUP<sup>®</sup>

Over the same period of time, there has been a very widespread perception that crime is getting worse.<sup>58</sup>

### Perceptions of Crime in the U.S., 1989-2014

Is there more crime in the U.S. than there was a year ago, or less?



#### GALLUP

The number of people who believe that crime is worse this year than last has hovered around 60–80 percent for decades, even as the number of people who have been the victim of a violent crime has fallen by a factor of three. Gallup goes so far as to say “perceptions of crime are still detached from reality ...federal crime statistics have not been highly relevant to the public’s crime perceptions in recent years.”<sup>59</sup>

How can this be? There is a wealth of data on crime in the United States, most of it freely available, and crime rate figures have been repeated endlessly in news stories. Surely this is an easily correctable misperception. (And it’s definitely a misperception. Although there are all sorts of issues in counting crime, violent crime rates are thought to be the most accurate type of crime data because the seriousness of incidents like homicide makes them harder to hide and easier to count.)

I don’t know for certain why perception is so far from reality in this case—I don’t think anyone really does—but the pattern fits what we’ve seen in experiments.

It was not until the 1970s that researchers investigated the human perception of statistical information in a serious way. Near the end of that decade, Hamill, Wilson, and Nisbett asked a simple question: How does statistical information change the perception of an anecdote?<sup>60</sup>

These researchers wanted to see if people would discount an extreme example when they were given statistics that showed it to be extreme. So they showed over a hundred people a *New Yorker* article about a welfare recipient:

The article provided a detailed description of the history and current life situation of a 43-year-old, obese, friendly, irresponsible, > ne'er-do-well woman who had lived in New York City for 16 years, the last 13 of which had been spent on welfare. The woman had emigrated from Puerto Rico after a brief, unhappy teenage marriage that produced three children. Her life in New York was an endless succession of common-law husbands, children at roughly 18-month intervals, and dependence on welfare. She and her family lived from day to day, > eating high-priced cuts of meat and playing the numbers on the days immediately after the welfare check arrived, and eating beans and borrowing money on the days preceding its arrival. Her dwelling was a decaying, malodorous apartment overrun with cockroaches ...<sup>61</sup>

This was a real person, but she was not a typical case, because almost no one stays on welfare for 13 years. One group of readers also saw statistical information showing this was so:

Statistics from the New York State Department of Welfare show that the average length of time on welfare for recipients between the ages of 40 and 55 is 2 years. Furthermore, 90 percent of these people are off the welfare rolls by the end of 4 years.<sup>62</sup>

The other group of readers was given false statistical information that made 13 years seem like a normal length of time:

Statistics from the New York State Department of Welfare show that the average length of time on welfare for recipients between the ages of 40 and 55 is 15 years. Furthermore, 90 percent of these people are off the welfare rolls by the end of 8 years.<sup>63</sup>

Then everyone was given a brief quiz with questions about their perception of welfare recipients such as:

How hard do people on welfare work to improve their situations? (1 = > not at all hard, 5 = extremely hard)<sup>64</sup>

As you might expect, most people came away from all of this with a rather negative impression of people on welfare—much more negative than a control group who did not read the story. But there was no meaningful difference in the opinions of those who read the real versus fake statistics, and no difference when the statistics were presented before versus after the story.

The description of the woman in her shabby apartment is so vivid, so real, so easy to connect to our own experiences and cultural stereotypes. It completely overwhelms the data. It's not that people didn't remember the average length of time someone stays on welfare; they were quizzed on that, too. The statistical information simply didn't figure into the way they formed their impressions.

I certainly don't blame readers for this; it's never worthwhile to blame your readers. Nor am I convinced I would be any different. I don't think it's clear enough that this woman was atypical, vivid examples are persuasive, and readers had no reason to be especially careful. Rather than shaking my faith in the intelligence of humanity, I just see this as a lesson in how to communicate better.

There have been other experiments in a similar vein, and they usually show that examples trump statistics when it comes to communication. In one study people were asked to imagine they were living with chest pain from angina and had to choose between two possible cures. They were told that the cure rate for balloon angioplasty was 50 percent and the cure rate for bypass surgery was 75 percent. They also read stories about people who underwent different surgeries. In some cases the surgery succeeded in curing their angina and in some it failed, but these examples contained no information that would be of use in choosing between the surgeries. Even so, people chose bypass surgery twice as often when the anecdotes favored it, completely ignoring the stated odds of a cure.<sup>65</sup>

Which brings us back to crime reporting. In major cities, not every murder makes the news. In different times and places the number of reported murders has varied between 30 percent and 70 percent of the total.<sup>66</sup> The crimes that get reported are always the most serious. Content analysis has shown that coverage is biased toward victims who are young, female, white, and famous, as well as crimes which are particularly gruesome or sexual. Yet these examples are the stuff from which our perceptions are formed. It's enough to make a media researcher weep:

Collectively, the findings indicate that news reporting follows the law of opposites—the characteristics of crimes, criminals, and victims represented in the media are in most respects the polar opposite of the pattern suggested by official crime statistics.<sup>67</sup>

Not only is crime reporting biased in a statistical sense, but the psychological dominance of examples means that readers end up believing almost the opposite of the truth. This is a type of media bias that is seldom discussed or criticized.

If you want the reader to walk away with a fair and representative idea of what the data means out in the world, then your examples should be *average*. They should be *typical*. This goes up against journalism's fascination with outliers. It's said that "man bites dog" is news, but "dog bites man" is not. But if we want to communicate what the bite data says we should consider going with "dog bites man" for our illustrative examples.

My favorite stories draw on both statistics and examples, using complementary types of information to build up a full and convincing picture. But generally, examples are more persuasive than statistics presented as numbers. Individual cases are much more relatable,



detailed, and vivid, and they will shape perception. The bad news is that poorly chosen examples can create or reinforce bad stereotypes. But this also means that well-chosen examples bring clarity, accuracy, and life to a story, as every storyteller knows.

## Who Is in the Data?

Data about people affects people's lives. Urban planners, entrepreneurs, social critics, police—all kinds of people use data-based representations of society in their work. This is why the issue of representation is so important. Changing how someone is perceived, or if they are perceived at all, can have enormous effects.

The “goodness” of a representation depends on what you want to do with it—the story you are telling—but in many cases it seems most fair to count each person equally. There is a nice alignment here between democracy and statistics, because the simplest way to generate data is to count each item in exactly the same way. Random samples are also very popular, but they are just a practical method to approximate this ideal. This moral-mathematical argument on the representativeness of data is almost never spelled out, but it's so deep in the way we think about data that we usually just say data is “representative” of some group of people when it approximates a simple count.

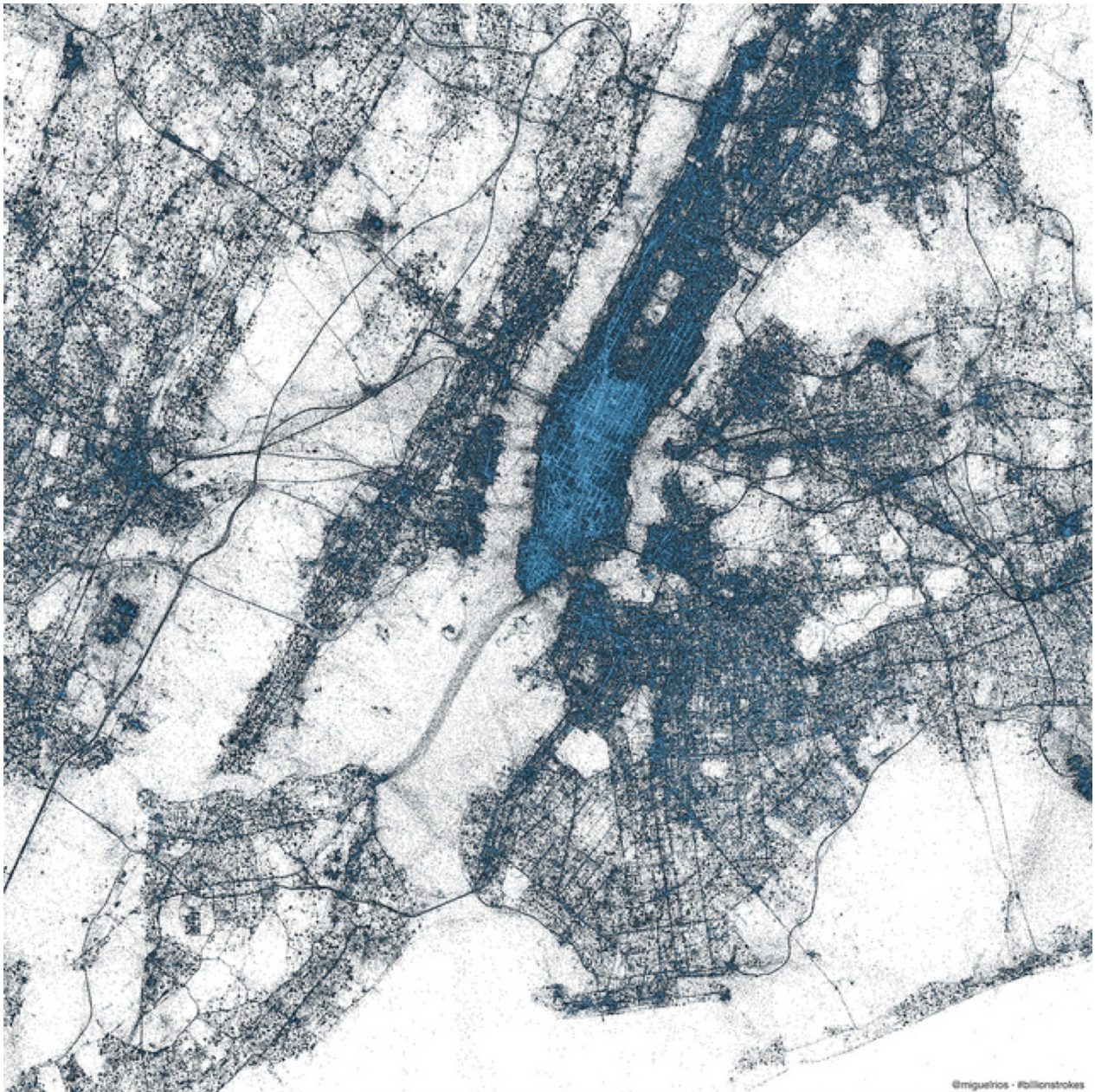
The data you have may deviate from this ideal in important ways.

Journalists have been trying to portray the public to itself for a long time. When you read an article about student debt that quotes a few students, these students are standing in for all students. Broadcast journalism's “person on the street” interview brings the reader into the story by presenting the opinions of people who are “just like them.” Of course, it never really works out that way; reporters only interview a small number of not-really-random people, and television crews tend to film whomever is easiest to get on camera.

When Osama bin Laden was killed in 2011, the Associated Press undertook a project to gather reactions from all over the world. Reporters rushed to pick up any camera they had and ask the same scripted question of many people. But which people? In practice it will depend on factors like which reporters are most keen on the project, who the reporters already know, who is easiest to get to, and who is most likely to speak a language the reporter understands. The project was meant to capture the global response to a historic event, but it's not clear whose voices are actually represented. A global, random video sample on a breaking news deadline would be quite a challenge, but perhaps you could try to get a certain range of country, age, race, gender, and so on.

Social media seems to offer a way out, because it represents so many more people. No doubt bulk social media analysis can be a huge improvement over a handful of awkwardly chosen sources. But social media isn't really representative either, not in the sense that a random sample is.

Here's New York City, as revealed by geocoded tweets:<sup>68</sup>



@miguelitos · #billionsofstrokes

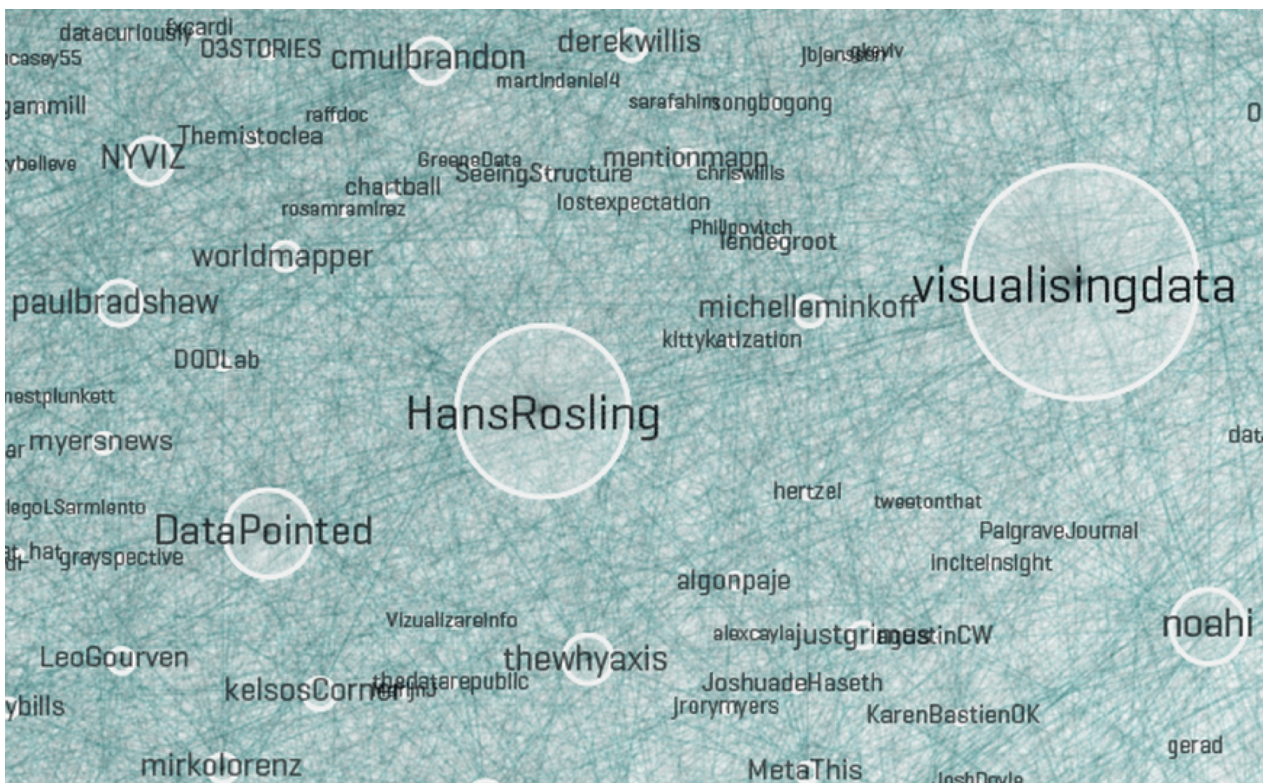
I find this map beautiful and revealing. It's not a map of geography or political boundaries, but a map of people. I love how it traces major transit routes, for example. But it is only a map of certain types of people, as I know from comparing it to a population-density map. There are large sparse areas in Brooklyn where plenty of people live, and Soho is definitely not as dense as Midtown. Also, only a few percent of tweets are geocoded. What sort of person uses this feature?

Not everyone is on Twitter, not everyone is Tweeting, and even fewer are speaking on the topic of your story. This data has a bias toward certain types of people, and you don't really know which kind of people those are. There is surely useful information to be got from social media, but it is not the same kind of information you can get from a random sample.



Whether or not this is a problem depends on your story. Twitter users tend to be affluent and urban, so if that's the population you want to hear from, you're good. If it's not, there may not be much to say from a Twitter analysis. Any representation of public sentiment created from social media data—a word cloud or anything else—will be biased in an unknown way. That is, the results will be skewed relative to a random sample, and the worst part is you won't know how skewed they are.

The way you choose your data can also create representativeness issues. Here's a visualization by Moritz Stefaner that is meant to show the “Vizosphere,” the people who make up the data visualization community.



Excerpt from the Vizosphere by Stefaner.<sup>69</sup>

Of course it's not really a visualization of everyone involved with visualization. To create this picture, Stefaner started with “a subjective selection of ‘seed accounts,’” meaning the Twitter handles of 18 people he knew to be involved in visualization. The 1,645 people included in the picture are all following or followed by at least five of these accounts.

The result is a very interesting representation of some people involved in visualization but certainly not everyone involved in visualization. Why these 18 accounts? Why not include people with four links instead of five? Part of the problem is that there is no universally accepted definition of who is “in” the visualization community, but even if there were, it's doubtful Twitter network analysis would be the way to find them all. This chart almost completely excludes the scientific visualization community, hundreds of people who have been doing visualization for decades.

Stefaner knows there are issues of this sort, and says so in the description of this image. There's nothing wrong with all this. But if it were to be presented as journalism, would readers need to parse the fine print to get an accurate understanding?

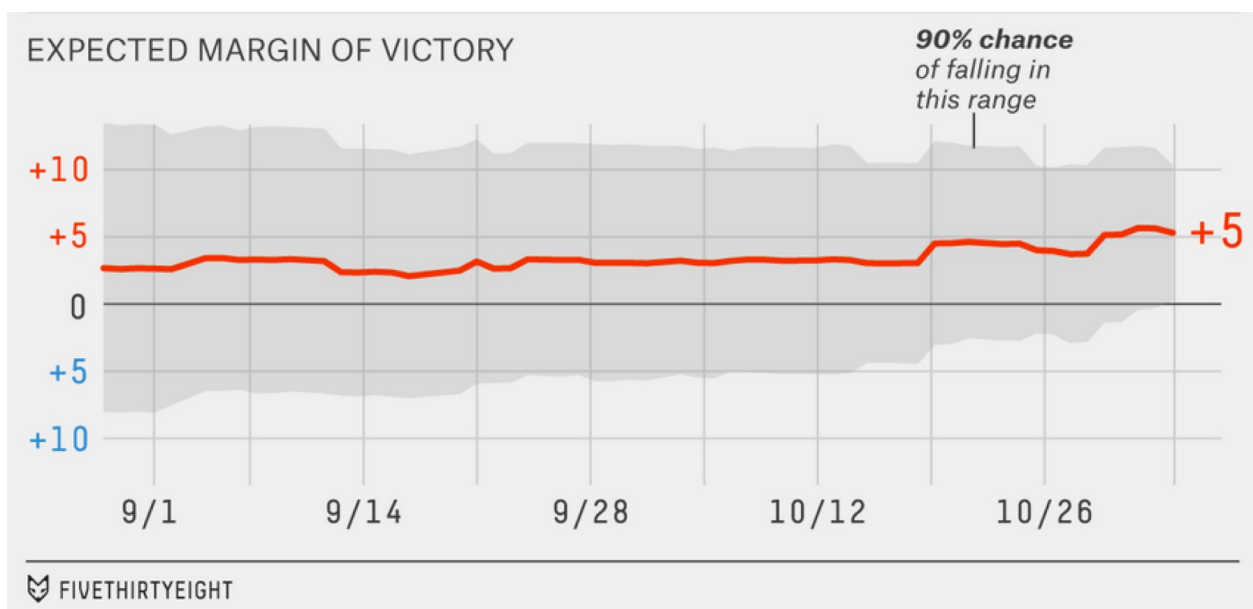
## Communicating Uncertainty

Uncertainty is a recurring theme in data work. It's familiar in a way, because we have all been unsure. But I don't think most people have a natural feel for quantitative measures of uncertainty. I suspect the best way to get a feel for uncertainty is to play with simulations of probabilistic things, but your readers won't have done that so we have to find other ways of communicating.

We've encountered quantified uncertainty many times already. The simplest way of presenting uncertainty is to give a range: 312  $\pm$ 7 miles. The margin of error of a sample is a more sophisticated measure that includes how often we expect the error to fall in that range: the poll numbers were 68 percent in favor, accurate to within 3 percent 19 times out of 20. Probabilities are also a kind of uncertainty: we analyzed the stoplight data and found that the odds were 2 to 1 in favor of the model with a working stoplight.

These sorts of numbers can be difficult to grasp on an intuitive level, yet the uncertainty in a result is a key part of that result. When the data is uncertain or leads to uncertain conclusions, it would be a lie to omit that uncertainty, or communicate it poorly.

There are many ways to communicate uncertainty. We can show it in a visualization by indicating the range of possible values.



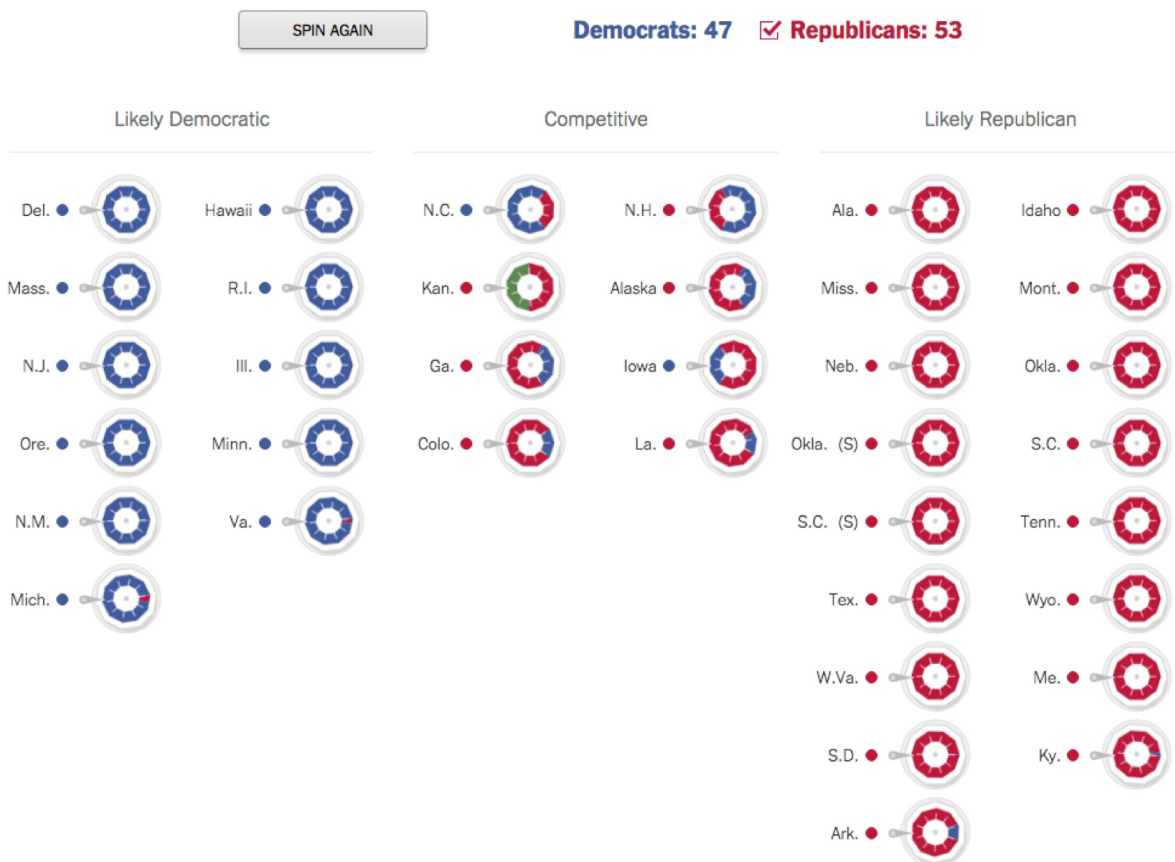
*Expected margin of victory in 2014 elections, from [fivethirtyeight.com](http://fivethirtyeight.com).<sup>70</sup>*

This image from the 2014 elections shows how the margin of error on the margin of victory changed over time.<sup>xxvi</sup> It clarifies something which is not otherwise obvious: The polls showed a consistent lead for months, yet it was only late in the race that victory was

particularly certain. All through September the odds were closer to 60/40, only narrowing substantially in the second half of October.

The gray region is the range of values where the outcome is expected to fall 90 percent of the time, the 90-percent confidence interval. The easiest way to compute this range is to simulate lots and lots of elections using a model that generates random outcomes according to the known uncertainty of the polling data, then find the 5th and 95th percentiles to cut off the outliers on the bottom and top. The 90 percent figure is arbitrary, really just convention, but it provides a reasonable balance. If we showed the entire 100 percent range of the data, the gray region would stretch to include every fluke scenario. If we showed only the central 50 percent then readers might come away with an overly narrow impression of the uncertainty, because the true result would fall outside the gray area half the time (assuming a properly calibrated prediction model).

We can also show uncertainty by presenting the results of simulations with randomness built in. *The New York Times* built a roulette machine to explain the uncertainties in its 2014 election predictions. Each state is represented by a wheel divided into colored segments according to the then-current probabilities that each party would win there. When the user clicks the spin button, all wheels spin and stop at random positions, producing a final tally of senate seats.



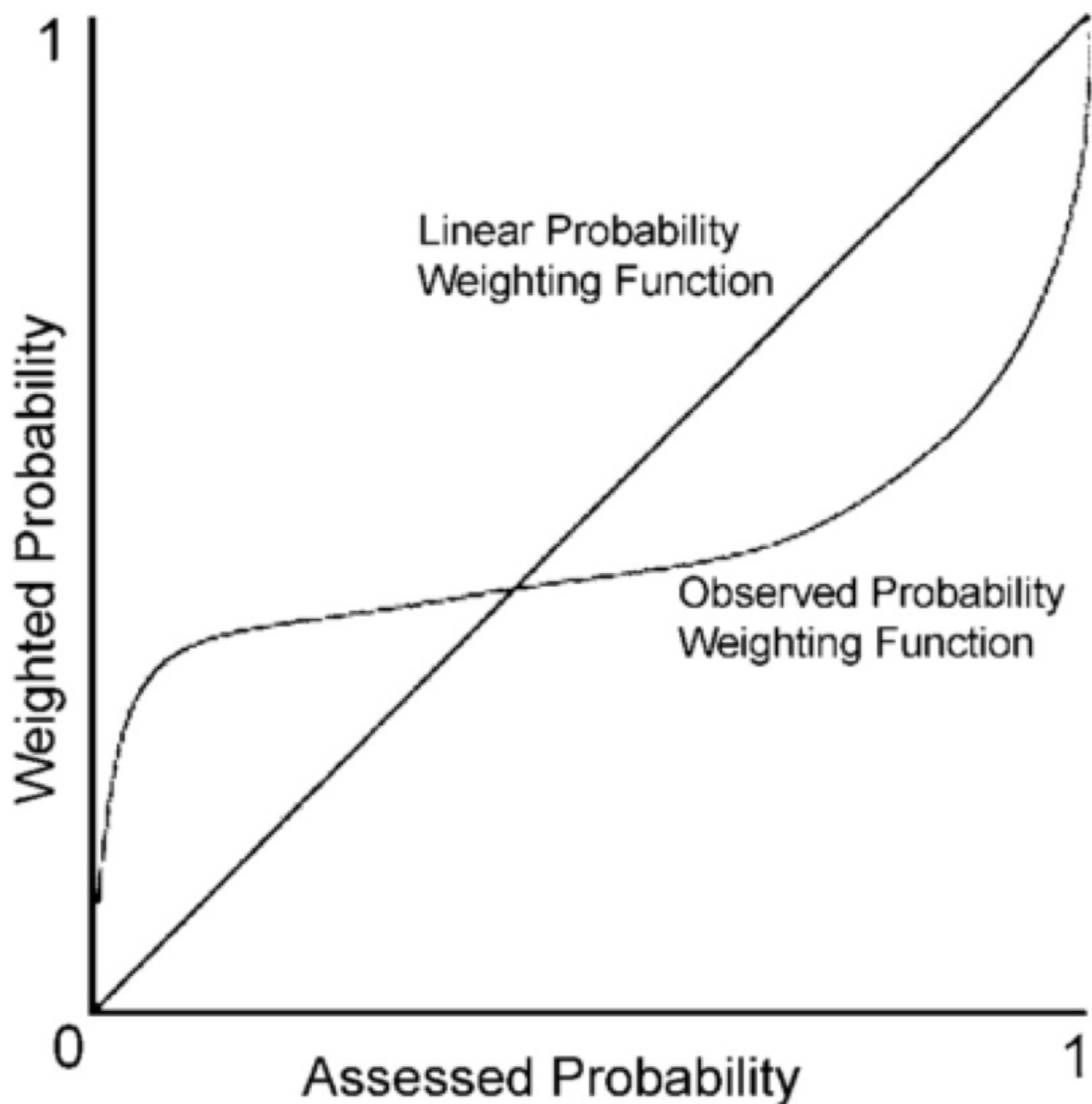
An illustration of the uncertainties in the outcome of the 2014 Senate races. Each time the user presses “spin again” the wheels rotate and stop at a random position. From *The New York Times*.<sup>71\*</sup>

This visualization relies on the same logic we used to analyze the stoplight data in the last chapter—it uses many simulation runs to show how the effects of chance shape the data we see. Understanding how some underlying reality leads to the observed data helps you figure out what the reality is when you are trying to interpret the data.

These examples both involve numbers with some probabilistic error. Sometimes what we need to communicate is just a probability by itself.

Humans have a nonlinear perception of numerical probabilities, as they do with many other perceptions (such as brightness which is perceived on a logarithmic scale). Daniel Kahneman and Amos Tversky pioneered the measurement of probability perception in the late 1970s with an experiment that gave people a choice between two bets with given odds and payoffs. They showed that people deviate in predictable ways from the best strategy of valuing a bet according to its average winnings, which you get by multiplying the probability of winning by the payoff. In these experiments, people acted as if small odds were much higher and large odds were much lower.<sup>72</sup> That is, people bet too much when the odds of winning were low, and too little when the odds of winning were high, even when they knew the exact odds!



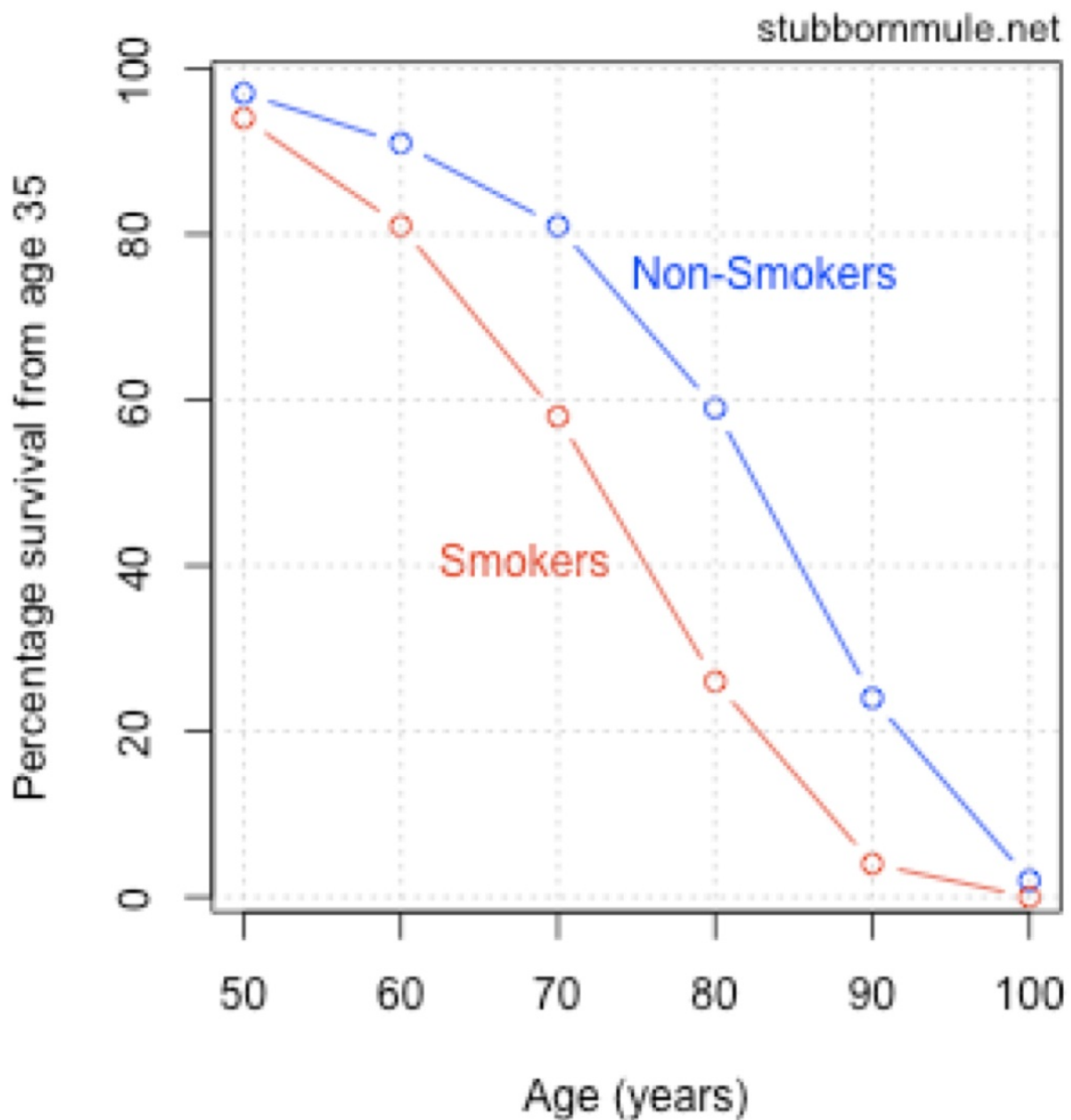


If this is how humans deal with probability figures generally, then we should expect people to exaggerate the probability of very rare events (like plane crashes) while underappreciating the probability of very likely events (like heart disease).

This is especially a problem when communicating small probability figures, such as rare risks. The probability of being struck by lightning in your lifetime is something around 0.0001.<sup>xxvii</sup> It's not immediately obvious what this means, but the chart above suggests that readers will tend to perceive getting struck by lightning as very much more likely than it actually is.

All sorts of things affect the perception of the probability of some event. If the event is very bad, we may perceive it as more common.<sup>73</sup> We will also imagine it to be more common if it's easy to bring examples to mind, a cognitive effect known as the *availability heuristic*. Thus, dying in a terrorist attack can seem just as probable as being struck by lightning even

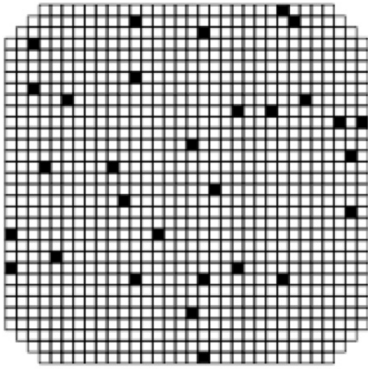




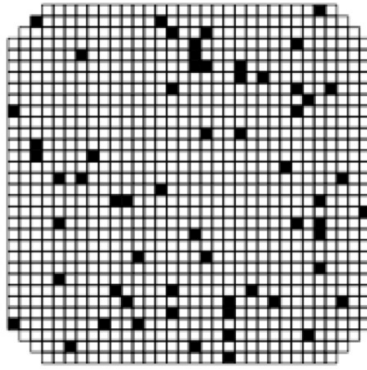
*Smokers versus non-smokers survival curves, from stubbornmule.net.*<sup>75</sup>

Everything you need to know is there, but it's a little hard to interpret. Let's see ...60 percent of non-smokers will live to 80 versus 25 percent of smokers. Figuring out what this data means requires far too much messing around with the chart and thinking through figures. Compare to the visualization:

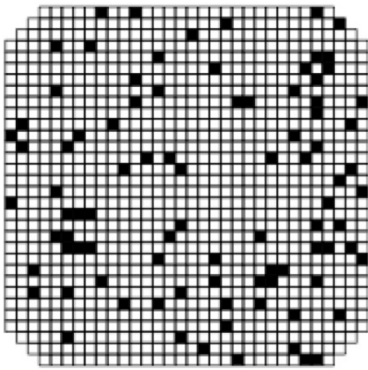
stubbornmule.net



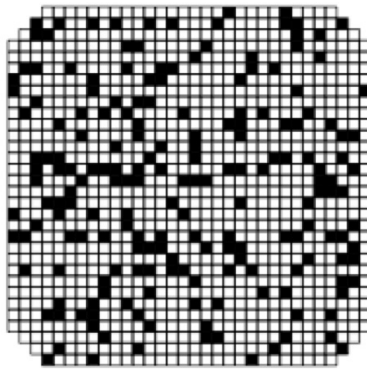
Non-smokers up to age 50



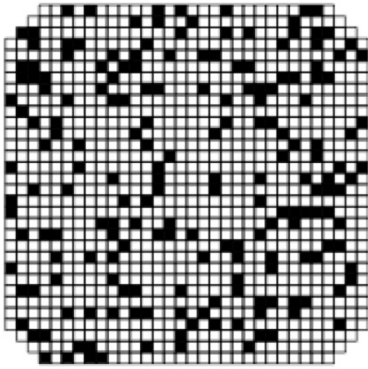
Smokers up to age 50



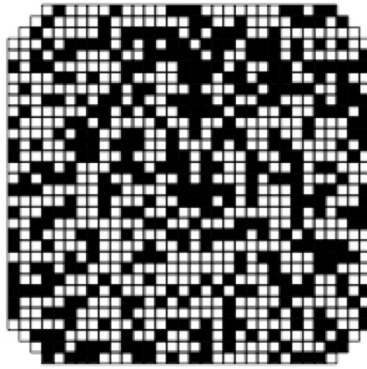
Non-smokers up to age 60



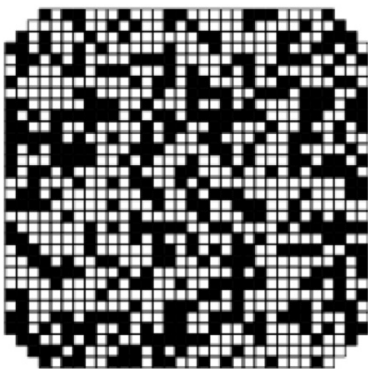
Smokers up to age 60



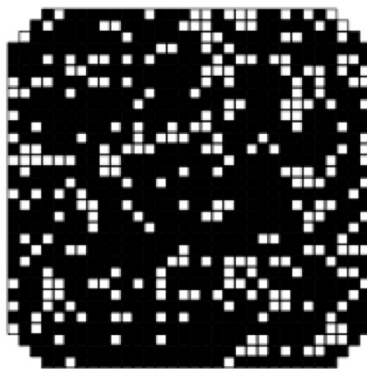
Non-smokers up to age 70



Smokers up to age 70



Non-smokers up to age 80



Smokers up to age 80

*Smokers versus non-smokers survival curves, from stubbornmule.net.*<sup>76</sup>

This visualization uses all the principles we've discussed. It represents probabilities as people, and compares probabilities both between smokers and non-smokers and between different ages. No one can know whether *they* will die from smoking, but visualizations like this can make the uncertainties personal.

There are lots of quantitative communication tricks and techniques you can pick up, and the visualizations here are not the last word in design. But the most important principle of communicating uncertainty is this: Communicate it. Don't let someone come away from your story with a warped sense of the risk, or too certain about something subtle. This is just basic respect for the reader and for the difficulties of knowing.

## Prediction

Prediction is important because action is important. What use is journalism that doesn't help you decide what to do? This requires knowledge of futures and consequences. Prediction also has close links to truth. Falsification is one of the strongest truth-finding methods, and it's prediction that allows us to compare our ideas with the world to see if they hold up. Prediction is at the core of hypothesis testing, and therefore at the core of science.

Journalists think about the future constantly, and sometimes publish their predictions: A particular candidate will win the election; the president will veto the bill if it's not revised; this war will last at least five years. It may be even more common to let sources make predictions: The analyst says that housing prices will continue to increase; a new study says this many people will be forced to move as the seas rise. Leaning on experts doesn't excuse the journalist from disseminating bad predictions unchallenged, and it turns out that experts quite often make bad predictions.

The landmark work here is Philip Tetlock's *Expert Political Judgment*.<sup>77</sup> Starting in 1984, Tetlock and his colleagues solicited 82,361 predictions from 285 people whose profession included "commenting or offering advice on political and economic trends." He asked very concrete questions that could be scored yes or no, questions like: "Will Gorbachev be ousted in a coup?" or "Will Quebec secede from Canada?"

The experts' accuracy, over 20 years of predictions and across many different topics, was consistently no better than guessing. As Tetlock put it, a "dart-throwing chimp" would do just as well. Our political, financial, and economic experts are, almost always, *just making it up* when it comes to the future.

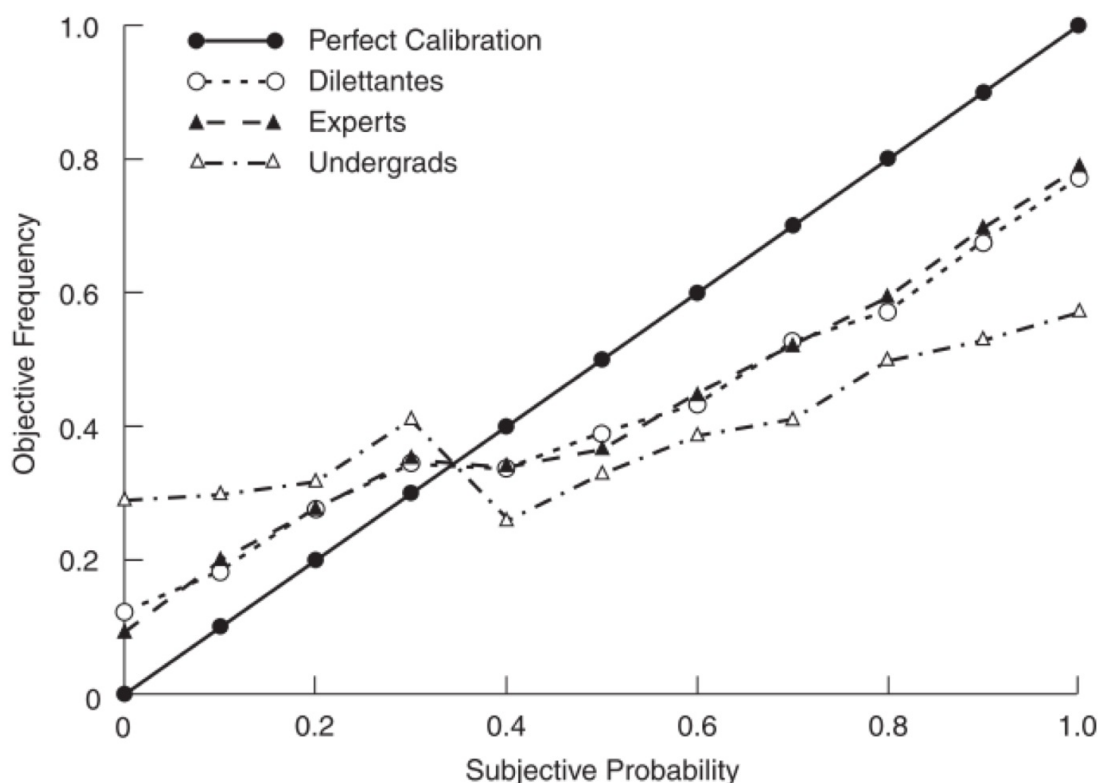
I suspect this is disappointing to a lot of people. Perhaps you find yourself immediately looking for explanations or rationalizations. Maybe Tetlock didn't ask the true experts, or the questions were too hard. Unfortunately the methodology seems solid, and there's certainly a lot of data to support it. The conclusion seems inescapable: We are all terrible at predicting our social and political future, and no amount of education or experience helps.

What does help is keeping track of your predictions. This is perhaps Tetlock's greatest contribution.

Although there is nothing odd about experts playing prominent roles in debates, it is odd to keep score, to track expert performance against explicit benchmarks of accuracy and rigor.<sup>78</sup>

The simplest way to do this is just to write down each prediction you make and, when the time comes, tally it as right or wrong. At the very least this will force you to be clear. Like a bet, the terms must be unambiguous from the outset.

A more sophisticated analysis takes into account both what you predict and how certain you think the outcome is. Out of all the predictions that you said were 70 percent certain, about 70 percent should come to pass. If you track both your predictions and your confidence, you can eventually produce a chart comparing your confidence to the reality. As Tetlock put it, "Observers are perfectly calibrated when there is precise correspondence between subjective and objective probabilities."



From Tetlock.<sup>79</sup>

Subjective probability is how confident someone said they were in their prediction, while the objective frequency is how often the predictions at that confidence level actually came true. In this data, when the experts gave something a 60 percent chance of occurring, their predictions came to pass 40 percent of the time. Overall, this chart shows the same general pattern found in other studies of probability perception: Rare events are perceived as much too likely, while common events are thought to be unduly rare. It also shows that expert knowledge helps, but only to a point. "Dilettantes" with only a casual interest in the topic did just as well as experts, and students who were given only three paragraphs of information were only slightly worse.

The overall lesson here is not that people are stupid, but that predicting the future is very hard and we tend to be overconfident. Another key line of research shows that statistical models are one of the best ways to improve our predictions.

In 1954 a clinical psychologist named Paul Meehl published a slim book titled *Clinical Versus Statistical Prediction*.<sup>80</sup> His topic was the prediction of human behavior: questions such as “what grades will this student get?” or “will this employee quit?” or “how long will this patient be in the hospital?” These sorts of questions have great practical significance; it is on the basis of such predictions that criminals are released on parole and scholarships are awarded to promising students.

Meehl pointed out that there were only two ways of combining information to make a prediction: human judgment or statistical models. Of course, it takes judgment to build a statistical model, and you can also turn human judgment into a number by asking questions such as “on a scale of 1–5, how seriously does this person take their homework?” But there must be some final method by which all available information is synthesized into a prediction, and that will either be done by a human or a mechanical process.

It turns out that simple statistical methods are almost always better than humans at combining information to predict behavior.

Sixty years ago, Meehl examined 19 studies comparing clinical and statistical prediction, and only one favored the trained psychologist over simple actuarial calculations.<sup>81</sup> This is even more impressive when you consider that the humans had access to all sorts of information not available to the statistical models, including in-depth interviews. Since then the evidence has only mounted in favor of statistics. More recently, a review of 136 studies comparing the two methods showed that statistical prediction was as good or better than clinical prediction about 90 percent of the time, and quite a lot better about 40 percent of the time. This holds across many different types of predictions including medicine, business, and criminal justice.<sup>82</sup>

This doesn't mean that statistical models do particularly well, just better than humans. Some things are very hard to predict, maybe most things, and simply guessing based on the overall odds can be as good (or as bad) as a thorough analysis of the current case. But to do this you have to know the odds, and humans aren't particularly good at intuitively collecting and using frequency information.

In fact the statistical models in question are usually simple formulas, nothing more than multiplying each input variable by some weight indicating its importance, then adding all variables together. In one study, college grades were predicted by just such a weighted sum of the student's high school grade percentile and their SAT score. The weights were computed by regression from the last few years of data, which makes this a straightforward extrapolation from the past to the future. Yet this formula did as well as professional



evaluators who had access to all the admission materials and conducted personal interviews with each student. The two prediction methods failed in different ways, and those differences could matter, but they had similarly mediocre average performance.

The idea that simplistic mechanical predictors match or beat expert human judgment has offended many people, and it's still not taken as seriously as perhaps it should be. But why should this be offensive? Meehl explained the result this way:

Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about \$17.00 worth; what do you think?" The clerk adds it up.<sup>83</sup>

Of course the statistical models used for prediction don't choose themselves. Someone has to imagine what factors might be relevant, and there is a great deal of expertise and work that goes into designing and calibrating a statistical model. Also, a model can always be surprised. An election prediction model will break down in the face of fraud, and an academic achievement model can't know what a death in the student's family will mean. Moreover, there can always be new insights into the workings of things that lead to better models. But generally, a validated model is more accurate than human guesses, even when the human has access to lot of additional data.

I think there are three lessons for journalism in all of this. First, prediction is really hard, and almost everyone who does it is doing no better than chance. Second, it pays to use the best available method of combining information, and that method is often simple statistical prediction. Third, if you really do care about making correct predictions, the very best thing you can do is track your accuracy.

Yet most journalists think little about accountability for their predictions, or the predictions they repeat. How many pundits throw out statements about what Congress will or won't do? How many financial reporters repeat analysts' guesses without ever checking which analysts are most often right? The future is very hard to know, but standards of journalistic accuracy apply to descriptions of the future at last as much as they apply to descriptions of the present, if not more so. In the case of predictions it's especially important to be clear about uncertainty, about the limitations of what can be known.

I believe that journalism should help people to act, and that requires taking prediction seriously.

## Going Further

You are probably no closer to finishing your next data project after reading this book.

I am painfully aware that the theory in this book is somewhat removed from the daily work of data journalism. You're going to need practical skills like working with spreadsheets, cleaning data, coding up visualizations, and asking civil servants for explanations. I've covered none of this craft.

Yet all of this work is guided by old and deep principles. Journalists are latecomers to quantitative thinking. That's unfortunate, because numbers can bring us closer to the truth. But only sometimes. Hopefully you now have a better sense of the limitations of data, and the ways we analyze and communicate data.

There's a lot more to learn.

There are an endless number of technical concepts relevant to data work. I've tried to give an authentic taste of the state of the art, and Bayesian statistics and cognitive biases are at the forefront of contemporary practice across many fields. Still, these presentations do not have the depth and detail needed to do real work; no one is going to learn to do statistical analysis from what I've written. Not exactly.

The good news is you don't have to learn everything at once. An education in statistics will give you powerful fundamentals that can be used to reason about subtle problems, but you won't need to do that every day. Also, that's what collaborators and mentors are for. A journalist's primary responsibility is to the story, and technical mastery comes from the experience of many solved problems.

It's not knowing everything that makes a technical professional, it's being willing to find out. I've used standard mathematical language in an effort to help you find more information; with a search engine, knowing the true name of something gives you the ability to summon it at will. So don't be surprised when you don't know something. If you're anything like me you'll get the code wrong the first time, even when you do know what you're doing. But never doubt that there is a logic underlying every equation and every line of code. These things are not magic; though the symbolic languages of data can be intimidating, there is nothing occult here.

My advice is to look always for the underlying sense of the thing, the plain-language explanation. This sense can be hard to find. When you ask a question like "why does a survey have a bell-shaped error distribution?" you will soon find yourself lost in inscrutable

proofs, answers that seem to presuppose you already know, explanations that don't really explain. This is an unfortunate comment on the state of our educational materials, but don't lose hope! Keep searching until you find an answer that makes sense.

Yet a technician is not a journalist. What will you be able to do with all of this understanding and ability?

Like any medium, it can take a while to find your voice in data journalism. Sure, you can do analysis and visualization and all the rest of it—but what are you saying? What questions are you asking? What is it that is so important, so urgent, that you must command a stranger's time to tell it to them?

I don't know of any way to discover what you want to say other than saying it. Just write. And report and code and visualize, but whatever else you do, put your work into the world. Then do the next one. As Steve Jobs said, real artists ship.

If you continue your study of the deep workings of data, you will discover entire worlds. You will retrace thousands of years of inspired ideas, re-experiencing each little epiphany as your own. You will gradually arrive at one of the most exciting frontiers of human thought, and you will join professionals in many other fields who are transforming their work through data. Quantitative ideas now pervade every aspect of the functioning of society, from health to finance to politics. It's impossible to understand the modern world without understanding data.

And if you do understand data, you will begin to see stories that others literally cannot imagine. We need those stories told. That is, perhaps, the best possible argument for learning more.

## Footnotes

- <sup>i</sup> You might as well expand that to the relationship between story and science. It's a vexing question. See, for example, Gelman and Basbøll.<sup>84</sup>
- <sup>ii</sup> The classic discussion of the human creation of categories is *Sorting Things Out: Classification and Its Consequences*.<sup>85</sup>
- <sup>iii</sup> For a thorough discussion of race on the census, see Snipp.<sup>86</sup>
- <sup>iv</sup> For a fantastic list of 20 reasons why quantification is difficult in psychology, see Meehl.<sup>87</sup>
- <sup>v</sup> For a really excellent exposition of the problems of counting “mass shooting,” see Watt.<sup>88</sup>
- <sup>vi</sup> Nehemiah 11:1.
- <sup>vii</sup> For more on these two unemployment surveys and the difference between them, see U.S. Bureau of Labor Statistics.<sup>89</sup>
- <sup>viii</sup> Actually 60,000 randomly chosen households, which is about 150,000 people. See U.S. Census Bureau.<sup>90</sup>
- <sup>ix</sup> Similar, but not identical, because Bernoulli initially considered sampling “with replacement,” where each person might be chosen more than once. This is probably because sampling with replacement is mathematically simpler, and Bernoulli worked with approximate formulas that become more accurate as the number of samples increases, rather than the very large numbers involved in calculating the number of possibilities directly, which require computers.
- <sup>x</sup> I'm indebted to Mark Hansen for the phrasing of these two key sentences.
- <sup>xi</sup> Before I get hate mail: Yes, it is wrong to say that there is a 90 percent chance that the true value falls within a 90-percent confidence interval. The contortions of frequentist statistics require us to say instead that our method of constructing the confidence interval will include the true value for 90 percent of the possible samples, but we don't know anything at all about this particular sample. The distinction is subtle but real. It's also usually irrelevant for this type of sampling margin of error computation, where the confidence interval is numerically very close to the Bayesian credible interval, which actually does contain the true value with 90 percent probability. See e.g. Vanderplas.<sup>91</sup>

**xii** Whether or not anything is “truly” random is a metaphysical question. Perhaps the universe is fully deterministic and everything is fated in advance. Or perhaps more data or better knowledge would reveal subtle connections. But from a practical point of view, we only care if these fluctuations are random to us. Randomness, chance, noise: There is always something in the data which follows no discernable pattern, caused by factors we cannot explain. This doesn't mean that these factors are unexplainable. There may be trends or patterns we aren't seeing, or additional data that might be used to explain what looks like chance. For example, we might one day discover that the number of assaults is driven by the weather. But until we discover this relationship, we have no ability to predict or explain the variations in the assault rate so we have little choice but to treat them as random.

**xiii** For a fantastic history of these ideas, see Ian Hacking's *The Emergence of Probability*.<sup>92</sup>

**xiv** Although the mathematics turn out the same, there's a useful distinction between something which we must treat as random because we don't know the correct answer (epistemic uncertainty) and something which has intrinsic randomness in its future course (aleatory uncertainty). The difference is important in risk management, where our uncertainty might be reduced if we did more research, or we might be up against fundamental limits of prediction.

**xv** Peirce's simple argument assumes complete statistical independence between the positions of every stroke in a signature. That's dubious, because if you move one letter while signing, the rest of the letters will probably have to move too. A more careful analysis<sup>93</sup> shows that an exact signature match is much more likely than one in  $5^{30}$  but still phenomenally unlikely to happen by chance.

**xvi** For a baggage-free introduction to applied Bayesian stats I recommend McElreath's *Statistical Rethinking*, or his marvelous lecture videos.<sup>94</sup>

**xvii** I'm referencing the butterfly effect, the idea that the disturbances from a butterfly flapping its wings might eventually become a massive hurricane. More generally, this is the idea that small perturbations are routinely magnified into huge changes. The early chaos theorist Edward Lorenz came up with the butterfly analogy while studying weather prediction in the early 1960s. In practice, this uncertainty amplification effect means there will be random variations in our data, due to specific unrepeatable circumstances, that we cannot ever hope to understand.

**xviii** This type of independent events model is also called a *Poisson distribution*, after the French mathematician Siméon Denis Poisson, who first worked through the math in the 1830s. But the nice thing about using a simulation of our intersection is that it's not

necessary to know the mathematical formula for the Poisson distribution. Simply flipping independent coins gives the same result. Simulation is a revolutionary way to do statistics because it so often turns difficult mathematics into easy code.

**xix** Maybe both of your hypotheses are wrong, and something else entirely happened. Maybe your models, which are pieces of code, aren't good representations of your hypotheses, which are ideas expressed in language. Maybe your data is the result of both a working stoplight and some amount of luck. Maybe the intersection was rebuilt after the second year with wider lanes and a new stoplight, and it's really the wider lanes that caused the change. Maybe the bureaucracy that collects this data changed the definition of "accident" to exclude smaller collisions. Or maybe you added up the numbers wrong.

**xx** Unemployment versus investment chart from Mankiw.<sup>95</sup>

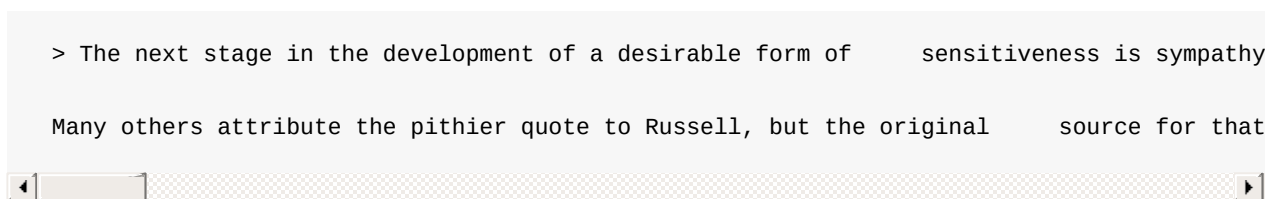
**xxi** But sometimes it *is* possible to tell which of two variables is the cause and which is the effect just from the data, by exploiting the fact that noise in the cause shows up in the effect but not vice versa. See Mooij et al.<sup>96</sup>

**xxii** Michael Keller, private communication.

**xxiii** I found this circulating on the Internet, and was unable to figure out who made it. Much love to the unknown creator.

**xxiv** It probably wasn't Bertrand Russell who first said, "The mark of a civilized human is the ability to look at a column of numbers, and weep." But per <http://quoteinvestigator.com/2013/02/20/moved-by-stats/> there is a history of quoting and misquoting a similar phrase. The original text is Russell's *The Aims of Education*:

```
> The next stage in the development of a desirable form of      sensitiveness is sympathy
Many others attribute the pithier quote to Russell, but the original      source for that
```



**xxv** I'll use *reader* as a generic name for the consumer of a story, with apologies to reporters working in other formats.

**xxvi** Totally fun to say.

**xxvii** Lifetime odds of being struck by lightning estimated at 1 in 12,000 by NOAA, based on 2004–2013 averages.<sup>97</sup>

## Citations

1. Denise Schmandt-Besserat, "Tokens and Writing: The Cognitive Development," *SCRIPTA* (2009): 145–154, [http://sites.utexas.edu/dsb/files/2014/01/TokensWriting\\_the\\_Cognitive\\_Development.pdf](http://sites.utexas.edu/dsb/files/2014/01/TokensWriting_the_Cognitive_Development.pdf).
2. "Table A-15: Alternative Measures of Labor Underutilization," U.S. Bureau of Labor Statistics, <http://www.bls.gov/news.release/empsit.t15.htm>.
3. Jonathan Stray, "Ethics in Data Journalism: Margin of Error in Bureau of Labor Statistics Reports," *Data Driven Journalism*, 15 January 2016, [http://datadrivenjournalism.net/news\\_and\\_analysis/ethics\\_in\\_data\\_journalism\\_margin\\_of\\_error\\_in\\_bureau\\_of\\_labor\\_statistics\\_rep](http://datadrivenjournalism.net/news_and_analysis/ethics_in_data_journalism_margin_of_error_in_bureau_of_labor_statistics_rep).
4. George Cobb, "The Introductory Statistics Course: a Ptolemaic Curriculum," *Technology Innovations in Statistics Education*, 1 (2007), <http://escholarship.org/uc/item/6hb3k0nz>.
5. James C. Scott, *Seeing Like a State* (New Haven: Yale University Press, 1998).
6. David Hestenes, "Oersted Medal Lecture 2002: Reforming the Mathematical Language of Physics," *American Journal of Physics*, 104 (2003), <http://dx.doi.org/10.1119/1.1522700>.
7. Brian Gratton and Myron P. Guttman, "Hispanics in the United States 1850–1990," *Historical Methods*, 3 (2000), <http://www.latinamericanstudies.org/immigration/Hispanics-US-1850-1990.pdf>.
8. David Niose, "Anti-Intellectualism Is Killing America," *Psychology Today*, 23 June 2015, <https://www.psychologytoday.com/blog/our-humanity-naturally/201506/anti-intellectualism-is-killing-america>.
9. G. Kitson Clark, *The Making of Victorian England* (New York: Routledge, 1962).
- 10.
11. Chris Davis and Matthew Doig, "State Scraps Felon Voter List," *Sarasota Herald-Tribune*, 12 July 2004, <http://www.heraldtribune.com/article/20040712/NEWS/>
- 12.
13. Matt Waite, "Handling Data About Race and Ethnicity," *OpenNews Source*, 20 June 2014, <https://source.opennews.org/en-US/learning/handling-data-about-race-and-ethnicity/>.

14. "Sixteenth Decennial Census of the United States, Instructions to Enumerators, Population and Agriculture," U.S. Census Bureau, 1940, <http://www.census.gov/history/pdf/1940instructions.pdf>.
15. Jens Manuel Krogstad and Mark Hugo Lopez, "'Mexican,' 'Hispanic,' 'Latin American' Top List of Race Write-ins on the 2010 Census," Pew Research Center, 4 April 2014, <http://www.pewresearch.org/fact-tank/2014/04/04/mexican-hispanic-and-latin-american-top-list-of-race-write-ins-on-the-2010-census/>.
16. "Directive No. 15 as Adopted on May 12, 1977," U.S. Census Bureau, 1977, <http://wonder.cdc.gov/wonder/help/populations/bridged-race/directive15.html>.
17. Jerzy Wojewoda et al., "Hysteretic Effects of Dry Friction: Modelling and Experimental Studies," *Philosophical Transactions of the Royal Society A*, 1866 (2008), <http://rsta.royalsocietypublishing.org/content/366/1866/747>.
18. "Employment Situation Technical Note," U.S. Bureau of Labor Statistics, 2015, <http://www.bls.gov/news.release/empsit.tn.htm>.
19. Neil Irwin and Kevin Quealy, "How Not to Be Misled by the Jobs Report," *The New York Times*, 1 May 2013, [http://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html?\\_r=0](http://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html?_r=0).
20. "How the Government Measures Unemployment," U.S. Bureau of Labor Statistics, 2015, [http://www.bls.gov/cps/cps\\_htgm.htm](http://www.bls.gov/cps/cps_htgm.htm).
21. "Employment Situation Technical Note."
22. Marianne Durand and Philippe Flajolet, "Loglog Counting of Large Cardinalities," in *ESA* (2003), 605–617, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.2718>.
23. Sir Arthur Conan Doyle, "The Adventure of the Blanched Soldier," in *The Case-Book of Sherlock Holmes* (1927), 54.
24. Mike Bostock et al., "One Report, Diverging Perspectives," *The New York Times*, 5 October 2012, <http://www.nytimes.com/interactive/2012/10/05/business/economy/one-report-diverging-perspectives.html>.
25. James Fallows, "Why to Get More Than 1 Newspaper, iPad Edition," *The Atlantic*, 22 October 2013, <http://www.theatlantic.com/national/archive/2013/10/why-to-get-more-than-1-newspaper-ipad-edition/280772/>.
26. Kypros Kypri et al., "Effects of Restricting Pub Closing Times on Night-time Assaults in an Australian City," *Addiction*, 2 (2011), <http://onlinelibrary.wiley.com/enhanced/doi/10.1111/j.1360-0443.2010.03125.x/>.



27. Ibid.
28. Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't* (New York: Penguin, 2012), 484.
29. Ibid.
30. Sanjoy Mahajan, *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving* (Cambridge: MIT Press, 2010).
31. Meier and Zabell, "Benjamin Peirce and the Howland Will."
32. Ian Hacking, "Telepathy: Origins of Randomization in Experimental Design," *Isis*, 3 (1998), <http://www.jstor.org/stable/234674>.
33. Gerard E. Dalal, "Why P=0.05?" <http://www.jerrydallal.com/LHSP/p05.htm>.
34. Anders Hald, "On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares," *Statistical Science*, 2 (1999), <http://www.jstor.org/stable/2676741>.
35. Robert Kass and Adrian Raftery, "Bayes Factors," *Journal of the American Statistical Association*, 430 (1995), <http://www.jstor.org/stable/2291091>.
36. Sharon Bertsch McGrayne, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy* (New Haven: Yale University Press, 2011).
37. Steven Raphael and Jens Ludwig, *Evaluating Gun Policy: Effects on Crime and Violence* (Chicago: Brookings Institution Press, 2003), 251–277, [http://home.uchicago.edu/ludwigj/papers/Exile\\_chapter\\_2003.pdf](http://home.uchicago.edu/ludwigj/papers/Exile_chapter_2003.pdf).
38. Ibid.
39. Steven D. Levitt, "Understanding Why Crime Fell in the 1990s: Four Factors That Explain the Decline and Six That Do Not," *The Journal of Economic Perspectives*, 1 (2004).
40. Raphael and Ludwig, *Evaluating Gun Policy: Effects on Crime and Violence*.
41. Kypri et al., "Effects of Restricting Pub Closing Times on Nighttime Assaults in an Australian City."
42. Raphael and Ludwig, *Evaluating Gun Policy: Effects on Crime and Violence*.
43. *Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970–1972* (London: Her Majesty's Stationery Office, 1978), <http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html>.

44. Franz H. Messerli, "Chocolate Consumption, Cognitive Function, and Nobel Laureates," *New England Journal of Medicine* (2012): 1562–1564.
45. Greg Mankiw, "A Striking Scatterplot," 29 March 2011, <http://gregmankiw.blogspot.com/2011/03/striking-scatterplot.html>.
46. Ibid.
47. Christian Rudder, "Exactly What to Say in a First Message," OKCupid blog, 2009, <http://blog.okcupid.com/index.php/online-dating-advice-exactly-what-to-say-in-a-first-message/>.
48. Milberger et al, "Tobacco Manufacturers' Defence Against Plaintiffs' Claims of Cancer Causation: Throwing Mud at the Wall and Hoping Some of It Will Stick," *Tobacco Control* (December 2006): iv17–iv26, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563590/>.
49. Andrew Gelman, "Statistics for Cigarette Sellers," *Chance*, 3 (2012), <http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics4.pdf>.
50. Bikaramjit Mann and Evan Wood, "Confounding in Observational Studies Explained," *The Open Epidemiology Journal* (2012), <http://benthamopen.com/contents/pdf/TOEPIJ/TOEPIJ-5-18.pdf>.
51. James F. Pagel, Natalie Forister, and Carol Kwiatkowi, "Adolescent Sleep Disturbance and School Performance: The Confounding Variable of Socioeconomics," *Journal of Clinical Sleep Medicine*, 1 (2007).
52. Judea Pearl, *Causality: Models, Reasoning, and Inference*, 2nd Edition (Cambridge: Cambridge University Press, 2009).
53. Danial Kaplan, *Statistical Modeling: A Fresh Approach*, Second Edition (ProjectMosaic, 2012).
54. John Stuart Mill, *A System of Logic*, Vol. 1 (: 1843), 455.
55. Matt Apuzzo and Adam Goldman, "Documents Show NY Police Watched Devout Muslims," *Associated Press*, 6 September 2011, <http://www.ap.org/Content/AP-In-The-News/2011/Documents-show-NY-police-watched-devout-Muslims>.
56. Philip Kitcher, *The Advancement of Science: Science Without Legend, Objectivity Without Illusions* (Oxford: Oxford University Press, 1993).
57. Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2013).

58. Jr. Richards J. Heuer, *The Psychology of Intelligence Analysis* (: CIA, 1999), <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/art11.html>.
59. Charles Sanders Peirce, "Some Consequences of Four Incapacities," *Journal of Speculative Philosophy* (1868): 140–157.
60. Tamara Munzner, "Visualization," in *Fundamentals of Computer Graphics, Third Edition*, ed. Peter Shirley and Steve Marschner (AK Peters, 2009), 675–707, <http://www.cs.ubc.ca/labs/imager/tr/2009/VisChapter/>.
61. Justin McCarthy, "Most Americans Still See Crime Up Over Last Year," Gallup, 21 November 2014, <http://www.gallup.com/poll/179546/americans-crime-last-year.aspx>.
62. Ibid.
63. Ruth Hamill, Timothy DeCamp Wilson, and Richard E. Nisbett, "Insensitivity to Sample Bias: Generalizing From Atypical Cases," *Journal of Personality and Social Psychology*, 4 (1980).
64. Ibid.
65. Ibid.
66. Ibid.
67. Ibid.
68. Angela Fagerlin, Catharine Wang, and Peter A. Ubel, "Reducing the Influence of Anecdotal Reasoning on People's Health Care Decisions: Is a Picture Worth a Thousand Statistics?" *Medical Decision Making*, 4 (2005).
69. Stray.
70. Jessica M. Pollak and Charis E. Kubrin, "Crime in the News: How Crimes, Offenders and Victims Are Portrayed in the Media," *Journal of Criminal Justice and Popular Culture*, 1 (2007).
71. Miguel Ríos, "The Geography of Tweets," Twitter, 31 May 2013, <https://blog.twitter.com/2013/the-geography-of-tweets>.
72. Moritz Stefaner, "The VIZoSPHERE, 2011," 2011, <http://www.visualizing.org/full-screen/29391>.
73. "Special Coverage of the 2014 Midterms," FiveThirtyEight, 4 November 2014, <http://fivethirtyeight.com/live-blog/special-coverage-the-2014-midterms/?#livepress-update-20137747>.

74. The New York Times, "Who Will Win the Senate?" 4 November 2014, <http://www.nytimes.com/newsgraphics/2014/senate-model/>.
75. Elke Weber, "From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events," *Psychological Bulletin*, 2 (1994).
76. Adam J. L. Harris and Adam Corner, "Communicating Environmental Risks: Clarifying the Severity Effect in Interpretations of Verbal Probability Expressions," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 6 (2011).
77. Ulrich Hoffrage et al., "Representation Facilitates Reasoning: What Natural Frequencies Are and What They Are Not," *Cognition* (2002), <http://www.sciencedirect.com/science/article/pii/S0010027702000501>.
78. "Visualizing Smoking Risk," *Stubborn Mule*, 21 October 2010, <http://www.stubbornmule.net/2010/10/visualizing-smoking-risk/>.
79. Ibid.
80. Phillip E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press, 2005).
81. Ibid.
82. Ibid.
83. Paul Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis: University of Minnesota, 1954).
84. Quinn McNemar, "Review of *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* by Paul E. Meehl," *The American Journal of Psychology*, 3 (September 1955).
85. William M. Grove et al., "Clinical Versus Mechanical Prediction: A Meta-analysis," *Psychological Assessment*, 1 (2000).
86. Paul Meehl, "Causes and Effects of My Disturbing Little Book," *Journal of Personality Assessment*, 3 (1986).