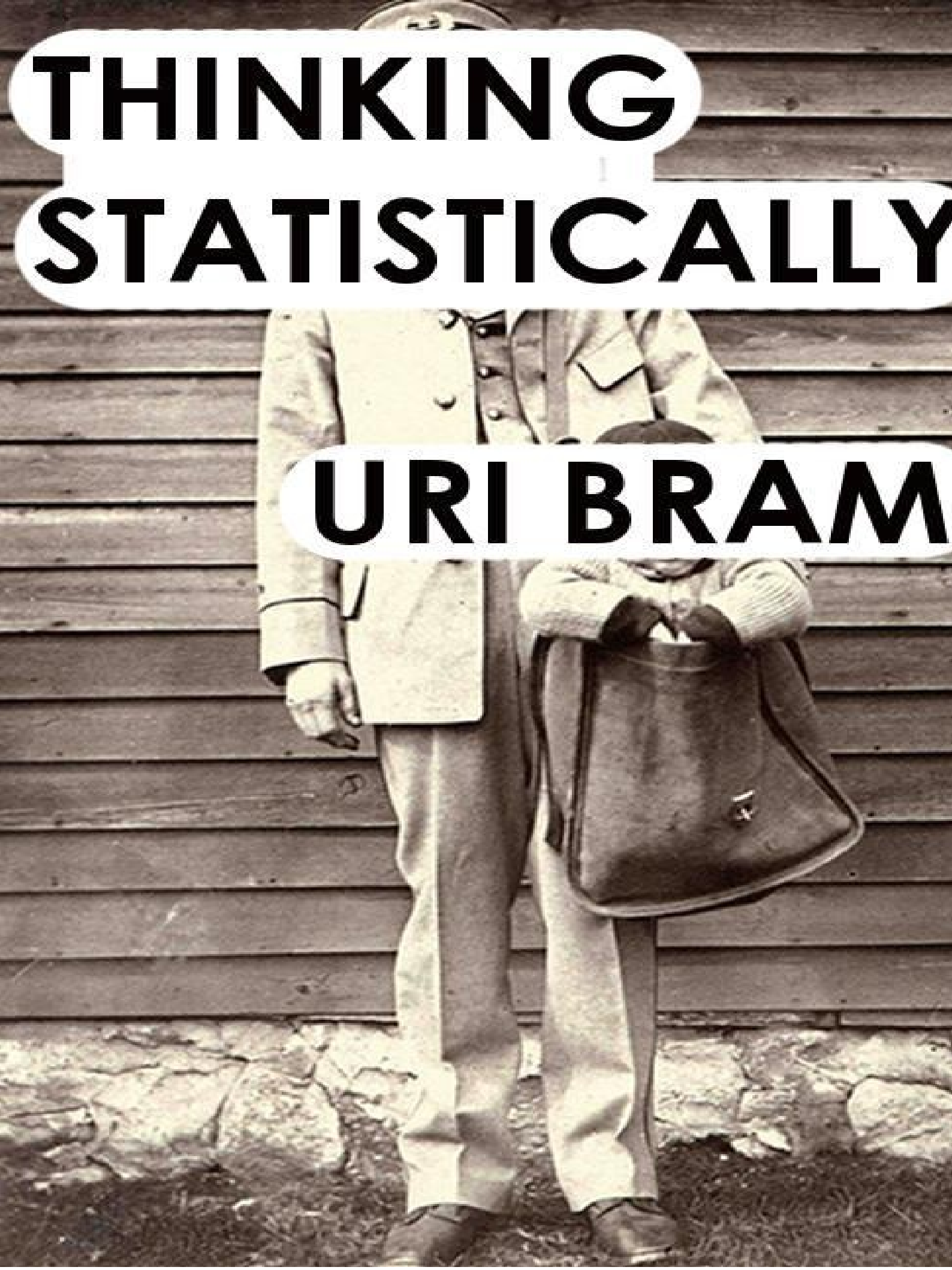
THINKING STATISTICALLY



URI BRAM

Thinking Statistically

by Uri Bram

to Elizabeth Swerdlow,
my favourite statistician.

TABLE OF CONTENTS

INTRODUCTION

CHAPTER ONE: SELECTION

- Buying apples
- Self-biasing data
- President Truman and the census
- The feedback effect
- Forgetting friends
- Congratulations

CHAPTER TWO: ENDOGENEITY

- Things cause other things
- A simple implicit equation
- Beware the error term, my child
- Omitted variable bias
- Why G.P.A. is stupid
- Causality loops
- People who switched
- Bill and Mark are endogeneous
- Consultants
- Social science and natural science
- Not all that correlates is caused
- Correlation, causation and endogeneity

CHAPTER THREE: BAYES

- Bayes matters
- Formula for Bayes
- Assimilating evidence
- Base rates and vaccinations
- Sally Clark
- Choosing jobs
- Sexuality
- Genius isn't everywhere
- Dating base rates

CODA

RECOMMENDED READING

SOURCES

ACKNOWLEDGMENTS

INTRODUCTION

It's not just you

You're smart. Really smart. People are always complimenting your curiosity and your gener knack for understanding things; maybe you even have a diploma from a fancy university, thoug that's irrelevant for our purposes. You're really smart, but (deep breath) you don't understand Baye Theorem. Or endogeneity. Or selection bias. Maybe you've never heard of them, maybe you rea about them once in a textbook. You definitely don't use them in your everyday life. It's not your faul for reasons completely beyond me, rarely does anyone explain these concepts to you properly unle you happen to be studying statistical techniques. But Bayes and selection bias and endogeneity a *concepts*, not techniques; understanding them is essentially independent from knowing how to find p-value, or the conditions for using a Student's t-distribution, or the proof of the Central Lim Theorem.

Concepts and abstractions are often useful even if you aren't exactly interested in the detai underneath them. It's good to understand that if the Central Bank raises interest rates then people ten to save more, without understanding the technical details of how a Central Bank effects interest ra changes. It's good to understand that there are sets of chords that harmonise together, and that man pop-songs are based on a few simple progressions, without understanding anything deeper abo music (or physics). It's good to understand the logic of evolutionary theory, even if you never plan t dust a fossil or examine mitochondrial D.N.A. Similarly, there's actually no good reason why yo have to learn statistical techniques in order to understand statistical concepts. The concepts on the own can be a huge help when examining and interpreting the world around us.

Most statisticians (and econometricians, and scientists) use statistics *informally* in their everyda lives, and a surprising number assume that everyone else does too. The world is made up information, and we make all kinds of decisions and judgments based on the information that reach us: if that information is systematically biased or incomplete then the decisions we make will b correspondingly wrong. Formal statistics is about taking precise, detailed information and provin exactly what that information can or can't tell us; informal statistics is about taking the vague, gener patterns of information that we see in everyday life and using the same basic statistical concepts make generally better decisions and judgments. Not perfect judgments, sure, and not ones you ca hang numbers on, but a better judgment is a better judgment and small improvements add up fast.

This book will teach you how to think like a statistician, without worrying about formal statistic techniques. Along the way we'll learn how selection bias can explain why your boss doesn't know h sucks (even when everyone else does); how to use Bayes' Theorem to decide if your partner cheating on you; and why Mark Zuckerberg should never be used as an example for anything. At n point will we do anything involving numbers. You can handle numbers just fine, if you want to, bu you don't right now.

This book is for you.

Alternative uses

If I'm going to be honest here — and if you and I are going to do a book together then honesty

very important — I have a second hope for this book, and then a third.

The second hope is that the book can be useful for people taking traditional statistics courses, as companion and refresher on the ideas behind the work they're doing. The truth is that nobody, r matter how statistically-minded they are, can consistently remember every idea they need to know ar the key intuitions behind how things work. Don't ever forget it — these ideas aren't easy, ar everyone has to look up the definitions sometimes. I hope that this book might provide a concise ar accessible summary of some of the key concepts to people knee-deep in formal tests.

The third hope is that —well— Hal Varian, the chief economist of Google, has said that "the sex job in the next 10 years will be statisticians... And I'm not kidding."* I hope this book will mal everyone who reads it as sexy as a statistician. Although people who choose to read this book seem t be exceptionally sexy already, so it might be hard to tell cause from effect.

CHAPTER ONE: SELECTION

Buying apples

So, it's Sunday morning (the birds in one another's arms, cats in the trees) and you're going to market stall to buy some apples. You pick up a couple; you poke them, shake them, and check ther for blemishes. If they seem pretty good then you grab another twenty and put the whole lot in yo shopping bag. You have just drawn an inference (that the apples are good) about a whole populatio (all the apples at this market stall) based on one small sample (the first few apples).

Whether you like it or not, you have just performed something called "inferential statistics." you were a more formal statistician you could run some fancy tests and figure out very precise thing about the apple-population: for example, the probability that all the apples at the stall are good, give that a certain number of randomly-selected samples were. But for any normal purposes, the first fe apples are a good-enough indicator of the quality of the rest.

They are only a reasonable indicator, though, if they are a (fairly) random selection from th population of apples as a whole. This is probably true if you chose them yourself. But what if, instea the stall-holder saw you gazing applewards and proffered you two to poke at? Now you would hav very little reason to trust that the sample-apples were representative of the general population: if th stall-holder is wily, and if she doesn't expect to see you again, then it's plenty likely that she uses h best apples as samples and so hides the fact that the rest are pretty bad. In statistics, this issue has th snazzy name of Selection Bias: basically, that your inferences will be biased if you use a non-rando sample and pretend that it's random.

Self-biasing data

In the apple example, the source of bias is obvious — the stall-holder's nefariousness. If her tric succeeds it's because we aren't paying attention; with a moment's pause, nobody would believe th the stall-holder's choice is even vaguely random. But often it's much harder to tell a good (informa sample from a bad one, partly because data samples can inadvertently bias themselves. There are lo of valid things we can say about a population from a smaller sample of it, *if* that sample is tru random. There are even lots of valid things we can say about a population if our sample is biased some systematic way that we understand and correct for. The real problems occur when our sample biased and we fail to account for that.

Imagine, for example, that you go to a daycare centre to take a survey of the children's ages. You sample is the children who showed up that morning, and your population is all the children who ev attend that daycare. So far so good: some children will be missing today, but assuming it's a norm morning there's no reason to think that the missing-ness will relate to age. Then you shout for all th one-year-olds to please stand up and be counted, then the same for all the two-year-olds, and final the three-year-olds. Any researcher would notice here that not all the children had stood. But a *bo* researcher would assume that the missing datapoints were missing at random, and could safely l ignored.

Now, any survey which can't absolutely coerce answers from every desired respondent wi
inevitably face some degree of non-responsiveness, and so long as the missing datapoints a
genuinely missing-at-random it really is safe to ignore them. However, in our daycare exampl
there's a clear possibility that the data is missing dependent on the value it would have taken. Younge
children are quite liable to be under-represented in our sample, since one-year-olds are less likely
understand our instructions, less likely to co-operate even when they do understand, and more likely
be distracted by a particularly interesting bug. Our sample is biased towards kids who've got the
heads together.

President Truman and the census

You might think that specific example is fairly silly, which admittedly it is. But it represents
genuine problem with sample selection that plagues even professional statisticians. Perhaps the mo
notorious example of selection bias in public surveys is Democrat Harry Truman's "upset" electio
victory in the 1948 U.S. Presidential election, which every pollster had predicted he was going to lo
by a landslide. Where did the pollsters go wrong? They had based their numbers on a series
telephone surveys — probably the most extensive, expensive and scientific telephone survey that ha
ever been done — and which confidently predicted a Republican victory. But who owned a telephon
in 1948? Well, generally wealthier people. Who happened to lean Republican.*

As another example, the U.S. Census Bureau has perpetual problems calculating the nation
population, despite a tremendous effort and the power of government behind them. And we're n
talking a couple of missing people: after the 2000 Census, the Bureau estimated that 4.5 millio
people were missing from its count, largely low-income Blacks and Hispanics.* For all kinds
reasons, the probability that a person will fail to return her census forms is related to her oth
attributes: her socio-economic status, her family situation, her location, her education levels, perha
her ethnicity. The people at the Census Bureau are fully aware of these issues, and make heroic effor
both to chase up non-responders and to publicise the shortcomings of their final counts. But the no
responses will inevitably occur, and occur in a strongly non-random way. This is not just an academ
problem: the census numbers are supposed to determine the funds that local governments receive fro
the federal government for social services. If residents of low-income urban neighbourhoods a
unusually likely to go missing from the census, low-income urban neighbourhoods will receive le
funds than they truly deserve.

The feedback effect

It's safe to assume that you aren't planning to run a major presidential poll or national census
the next few weeks, but similar effects haunt our everyday lives. For example, "feedback systems"
schools, companies, charities, and governments often suffer because the probability of a subordinate
opinion actually reaching the authorities is strongly related to whether the person's opinion wa
positive or negative. Bosses, donors, and academic deans often have impressions about the
organisations that seem utterly wacky to their less-pleased subordinates. But when the bosse
knowledge is based on the feedback that has been volunteered to them, and if they fail to correct f
the bias in this sample, it's little wonder that they reach incorrect inferences. In many wo
environments, where there are real or perceived dangers to expressing a negative opinion about th
boss' previous decisions, the kind of person who is most likely to volunteer her opinion is someon

who feels positively about the issue under discussion. Many bosses are aware of the issue in theor but it is surprisingly rare to find one who accounts for the effects of selection bias on the feedback sh herself receive—and even rarer to find one who corrects for it. In some ways, this explanation of wh superiors can be so delusional is rather comforting. It posits that many instances of supervisor madness are not the result of malice or idiocy, but rather genuine unawareness due to impeded data flows (and resultant sample selection bias).

If you find yourself in a supervisory position, you can outsmart and outperform your rivals b putting in place systems that *genuinely* ensure that the probability of feedback reaching you is n dependent on the values that feedback takes. There are many possible systems that could wor Anonymous comment-boxes are the right kind of idea: by disconnecting ideas from names, they sto subordinates from expecting reward or punishment depending on the value their feedback too However, there are often doubts about quite how anonymous an anonymous system is, and it difficult as a supervisor to weigh the negative views on slips of paper as strongly as the positive view you hear in person. As a result, the best system is probably one that involves encouraging negativ feedback in person: finding subordinates you trust and asking them what *other* people complain abo most; encouraging a "Devil's Advocate" mentality during projects and meetings; or askin subordinates what the least-best thing about a particular project is ("yes I know you think this proje is wonderful. But if you *were* to think it wasn't wonderful, what would you think wasn't wonderf about it?") Ultimately, your biggest strengths will be your simple awareness that negative feedbac must exist in your organisation, and your determination to keep looking for it despite th psychological temptation for positives to be reinforced.

When you're in the position of the supervised, rather than the supervisor, you can remind yourse of the importance of raising relevant issues with supervisors even when the issue seems obvious, wh you'd previously assumed that if nothing was being done it was because the higher-ups didn't care. the best case they may be willing to re-examine the issue once made aware of the rest of the evidenc but at the very least you can have the comfort of realising that you're not being treated wi purposeful malice. Understanding statistics can be an unexpected source of inner peace.

Forgetting friends

By contrast, the following advice may have a negative effect on your self esteem— sometime noticing sample selection bias can knock your sense of self importance. Just don't say I didn't wa you.

When I first got to America I noticed that a lot of people seemed to remember me long after I forgotten them. This led to all kinds of awkwardness, since they knew my name and where we'd m while I had no idea who they were. But it wasn't fair, I thought: of course they remembered me— was the guy with the curious accent, and they were just one of the hundred identical Americans (sor Americans) that I'd met that month.

I mentioned this odd phenomenon to a religious Jewish friend. "I have the same thing!," he said- "everyone remembers the guy with the Jewish head-covering, but how am I meant to distinguish a these identically-dressed hipster girls?" A short-haired female friend had a similar response: "I totall get you," she answered, "everyone remembers the short-haired girl, we're pretty rare here, but there no way I'm going to remember them as well. It's not fair, it's just not fair."

By this point I was was hugely suspicious that, in fact, all three of us had in fact over-estimate how memorable we were relative to other people: on reflection, all three of us had fallen victim sample selection bias. The problem was two different kinds of missing data.

First, in my own interactions, I lack a lot of data about the times when other people forget m
Everyone has (presumably) had moments where they couldn't remember someone's name fro
Adam, but managed to fake through a short conversation by just saying "you" or "buddy" or "heyyy
When that data has the value "I forgot your name, you remembered mine" then the data alwa
reaches me, in the sense that I'll always know whether I remembered your name or not. But there a
lots of instances when someone else forgets my name completely, spends a five-minute conversatio
calling me "buddy," and I just didn't notice anything. Round one to sample selection bias.

The second class of missing data involves the endless interactions that don't include me at all
which one party forgets the other party's name: where my religious Jewish friend, or my short-haire
female friend, forgets the name of someone who remembers them. Out of the total population
forgotten-name incidents (the ones where I forget someone's name; the ones where someone el
forgets mine; the ones where Person A forgets Person B's name; the ones where Person B forge
Person A's), the last group is almost completely invisible to me. Except for in very exceptiona
circumstances, I am never going to be aware of situations in which someone else forgets anoth
someone else's name—even though, intellectually, it's clear that this must (often) happen. Round tw
to sample selection bias.

The data that reaches me, then, is overwhelmingly skewed towards the instances where someon
remembers me and I don't remember them. Due to my biased sample, I wrongfully draw the inferen
that I'm unusually memorable (yes, yes, it's hard being a celebrity). Now, presumably some peop
forget names more and some people are forgotten more; it's possible that you, personally, really a
so unique that everyone remembers you while you can't remember them. However, it's important
realise that sample selection bias could trick you into feeling this way even when it isn't actually tru

Congratulations

There are actually many areas of personal life where selection bias occurs because we on
experience "first person" the things that happen to us, and don't always notice that we don't recei
the same stream of data from other people's first-person experiences.

When I was younger I learnt (very basic) guitar. I played mainly in my room, in the class
"awkward teenager croons melancholy love-songs" mold. Sometimes, feeling plucky, I even did Ope
Mic Nights. And invariably, after the show, people would come up to congratulate me and say ho
much they enjoyed my performance— I hate to brag but some even said I was the best in the who
show. The first kind of selection bias here is obvious: if someone thinks you played badly, they're n
too likely to tell you to your face. Almost-nobody comes up to a performer at a friendly open mic a
says "Hey!, I saw your performance and I just wanted to say that you're a mediocre singer and
barely-competent guitar-player," despite the fact that many open-mic guitarists are indeed medioc
and barely competent. Out of the entire population of opinions about my playing (positive opinio
and negative opinions), the sample that reaches me is strongly biased towards the positive ones.

The second kind of selection bias is a little more subtle: immediately after a performance, yo
tend to drift away from everyone else, and don't get to count how many times each of the oth
performers gets told how wonderful they were. Intellectually you know this is happening, b
pragmatically you forget about it. So out of the entire population of opinions about open-m
participants (the opinion that you were good; the opinion that you were bad; the opinion that someo
else was good; the opinion that someone else was bad) you'll hear a lot about how good you wer
almost-nothing about how bad you were, a decent amount about how bad other performers ("some
those guys shouldn't be on a stage at all, they're so barely competent") and maybe a little about ho

good some other performers were ("that girl just before you was amazing! I mean, I liked yours mo
obviously. But, y'know, she was ok.") The sample of opinions that reaches you is massively biase
and so you will greatly over-estimate how much other people liked your music and under-estima
how much they liked other performers'. Of course the other performers' biggest fans aren't talking
you right after the show - they've gone to talk to their own favourite musician. As a consequence, yo
over-estimate how good you were compared to other performers.

Constantly being aware of the extent that selection bias affects our self-perception would l
disastrously depressing (I did warn about that earlier, right?) Often it's probably better for us to igno
the sample selection bias, enjoy strumming on the big stage, and gracefully accept the compliment
afterwards. But if you get to the point where it feels like *everyone's* telling you you're an amazin
musician and you could *definitely* make it if you went professional, it's probably good to stop an
think about the selection bias in your feedback channels. *Before* you quit your day-job.

CHAPTER TWO: ENDOGENEITY

Things cause other things

"Endogeneity" is one of those words so thoroughly ignored that it doesn't even make it into mo normal spellcheckers. This is a minor outrage, because endogeneity is one of the most useful concep in all of statistics. It is also impossible to describe succinctly. The simple (and not necessarily helpfu definition is that something is endogeneous if it is determined *within* the system of interest. B contrast, something is "exogeneous" if it is determined by factors *outside* that system.

In biology, something endogenous originates within the organism, tissue or cell you're looking a something exogenous originates outside it. The same idea applies in social science (and in everyda life), except there is no perfect equivalent to cells and tissues and organisms. But social scientif arguments are composed of models, and real life is composed of informal, implicit models: when w think about whether one thing does or doesn't cause another, we are creating an implicit ment equation with "the things doing the causing" on the left-hand side and "the things that are caused" o the right. These models are, in fact, the relevant analogues for cells or organisms in biology, and if w construct the models poorly then we run into all kinds of problems. For example, perhaps our mod says that X causes Y when in fact Y causes X, or perhaps it claims that A causes B when in fact C causes both of them. Both of these are forms of endogeneity problem, for a precise reason that will b discussed later. For now, we can think of it like this: if our model claims that X causes Y, when in fac Y can cause X, then "the fact of Y causing X" is an important phenomenon that is occurring *outsi* the logic of our model (this one is a little complicated); if our model claims that A causes B when i fact C causes both of them then C is an important phenomenon from outside our model is influencin the outcome (this one is much simpler).

If these ideas sound familiar, perhaps that's because you've heard the famous maxim: "correlatio does not imply causation." Well spotted, cunning reader: correlations masquerading as causations ar indeed one type of endogeneity problem, and we'll discuss those later in the chapter. To understan endogeneity, though, we must first understand those informal, implicit mental models I mentioned. S here we go.

In real life, when we see an output of interest, we build a kind of model in our heads to expla how it occurred. The questions that we need to answer are basically, "Which inputs went in to caus this output?," and "How much influence did each of those inputs have?" The problem could b something like, "Why doesn't Little Mikey have any friends?" The output of interest is "Number friends that Mikey has," which in this case is zero, and the inputs could be things like "Little Mike keeps punching other children in the face," which has a very strong effect on the output, and "Litt Mikey has this really annoying way of sneezing," which does have some impact but not nearly much. The fact that Input A has a large negative effect on the Output (number of friends) implies th if we could reduce Input A, we would increase the Output accordingly; if we reduced Input B, whic has a smaller negative effect on the Output, then the Output would increase but not by as much.

As will be clear from the above paragraph, an implicit model can be written out in long-form, b it'll quickly get lost in the details and it's hard to convey the model with precision. Even in our simp model of Little Mikey's popularity—two inputs and one output—was hard to keep track of whe

written in prose. There is a solution, but it's often less popular than Little Mikey: equations. Whe
Stephen Hawking was writing *A Brief History of Time* (a life-changing read, by the way), he was to
that every equation he put in it would cost him half his readership. He limited himself to one and th
book sold 10 million copies, so if the warning was correct then that was a pretty expensive equatio
Unfortunately there's no way I can write the next section without using equations, and lots of them
They'll be nice ones, I promise; when it comes to equations I'm as bad as anyone, and when I s
them in a paper I tend to skip over as fast as my eyes will carry me. But since we all do use the
(implicitly) in our everyday lives, and since they never seem to bother us there, they can't possibly b
as scary as they look.

A simple implicit equation

Suppose you're standing in the supermarket and trying to decide which line to join. You see
couple of families with trolleys full of food, and a couple of lonely singles with a small basket eac
You make a quick mental calculation: each of the families will probably take 10 minutes to clear th
line, each of the singles will probably take 5. Congratulations! You've just created an implicit menta
equation. It looks like this:

$$\mathbf{10f + 5s = t}$$

f = # of families in line in front of you
s = # of singles in line in front of you
t = time you'll have to wait to reach the front of the line

That's not so bad, right? On the left of the equals sign we have some inputs, and on the right of th
equals sign we have an output, and we can predict the output using some function of the inputs. W
call the output the "dependent variable" because it can't vary freely within our model: its value
dependent on the values taken by the inputs. Which makes sense, because the whole point of o
model was to predict the value of the output (time spent in line) based on the values of the inpu
(number of families or singles in front of you). Similarly, we call the inputs "independent" becaus
their variation cannot be determined by any of the other variables in the model. Like teenager
independent variables won't let anyone tell them what to do or what to be.

Now, obviously there's something missing here: those "10 minutes" and "5 minutes" figures wer
only guesses, they weren't some kind of (impossibly precise) measure of exactly how long eac
person will take. We need to add something called an *error term*. The error term sweeps up all th
random variation and represents it as a single letter. We can now write the proper implicit equation –
again, you're doing something like this in your head every time you see a shopping line, whether yo
realise it or not.

$$\mathbf{10f + 5s = t + e}$$

f= # of families
s= # of singles
t = time you'll have to wait
e = the error term

This equation tells us that if there's 2 families and 3 singles in line in front of you, you can expe to wait (10 x 2) + (5 x 3) = 35 minutes, *plus-or-minus* the random error term. Few things in this wor can be predicted with absolute certainty, and we know that some of the families will only take minutes while others will take 12; some of the singles will only take 4 minutes and others 6. Oft those random errors will cancel out—if someone takes a minute more than expected, and someo else takes a minute less, then you're back on track overall—but if you're unlucky you'll get stuck in line where *everyone* is slower than expected. However, you know that the average total wait time wi be 35 minutes (assuming, of course, that your original equation was well-designed), and that a variation around that will be random. While there is some probability that the wait will be longe there is exactly the same probability that the wait will be shorter, and small differences from 3 minutes are much more probable than large ones. Now we have a proper estimate of the time we wait in line: 35 minutes, give or take something, but probably not much. Right?

Beware the error term, my child

At this point you may be sitting back smugly, wondering what all this error-related fuss is abou That was easy! But error terms are sneaky little monkeys, they start making problems when you lea expect it. In the shopping-line equation we designed, the error term was *uncorrelated* (as it should b with the independent variables — the random factors swept together in the error term were complete unrelated to the independent variables (the number of families, the number of singles) that we we using to make predictions. This was why we could use the formula to predict how long we would hav to wait in line—we knew that, while there would be some variation from the expected time, it wou be random variation and there would be no way to predict it from the information available to us.

We assume, by definition, that the error term is random variation — that errors are clustere around zero, that under-estimates are as probable as over-estimates, and that small errors are mo probable than large ones. A model with random error in it is still a useful model — not quite as goo as a perfect one, sure, but still worth a lot. Knowing that the shopping line will take somewhere ne 35 minutes is a very good thing, even if you also know that you could easily wait 5 minutes more less. But what happens if the error term secretly includes factors that are not, in fact, random? If the exists any variable that is correlated with the error term then the error term is non-random — this necessarily true, basically because of what it means to be truly random. This means that th differences between our predicted outcome and the true outcome will be systematic, not random: you had some data about the variable that was secretly correlated with the error term, you would the be able to predict in different situations whether the model would overestimate or underestimate th output.

But if our model is wrong systematically, not randomly, it is not very useful for makin predictions. Think again about our model of the shopping line; suppose we see a single line with singles in it. Our model claims that each single will take 5 minutes to clear the line, so in 20 minute time we'll get to the front—give or take some random error. However, in real life, this is not qui right—shoppers don't distribute themselves randomly into lines. Why are there 4 singles, one aft another, in a single line? Perhaps we are looking at the special "10 items or fewer" line, which is wh no families joined it. In this case, these specific singles will probably all take less than the 5 minu average, and our model will *systematically* overestimate how long we'll wait if we join this line.

As a rigorous definition, a variable is endogenous within a given system if it is correlated with th error term in the equation. Why? Because our equation is supposed to give us two things: an outp and an error term. Our output should be predicted by our inputs, and the error term should sweep u

all the sources of random variation that can make individual observations vary from that predicte value. The error term, then, should be determined by random factors completely outside the system - luck, the weather, rolls of the dice. If, instead, the error term is being (partly) determined by a variab inside the equation, our model's predictions will be systematically off. The variable in question called "endogenous"—from the Greek word *endon*, meaning 'within,' and possibly the Greek root *ge* meaning birth—because it is (illegally, improperly) helping to determine the error term from insid our system.

Omitted variable bias

One way to destroy a model through endogeneity is to leave out an important variable that needed to explain the outcome in question. Any part of your outcome that would have been explaine by this missing variable will have to be swept into the error term, which will now be non-random; will contain the entirety of the omitted variable. For example, the formula for the area of a square (of course):

height x width = Area

where height = width. But imagine that your model of a square was simply:

height = area + error

This is a monumentally stupid formula for the area of a square, but let's run with it for now. If yo believed in it you would measure a bunch of squares and notice that your formula did have some kin of truth. You'd see a strong correlation between height and area, and that the higher a square was th bigger its area tended to be. You'd predict that a square of height one should have area = 1, which correct; that a square of height two should have area = 2, when in fact it has area = 4; and that a squa of height three should have area = 3, when in fact it has area = 9. The error in your predictions completely non-random: the error gets greater as the height gets greater, for a start, because the erro "contains" the influence of the omitted variable "width." Furthermore, for any square with are greater than one, your prediction is always an under-estimate. If you gather a large set of squares the your error terms will not average out; your formula will systematically under-estimate the areas squares.

Why G.P.A. is stupid

A very important example of a model whose stated purpose is destroyed by omitted variable bia is college G.P.A. In the U.S., grades tend to be calculated on a 0-to-4 scale where an A-grade is wort 4.0, an A- is worth 3.7, a B+ is worth 3.3, and so forth. A student's overall G.P.A. is calculated as th average of her scores in all the courses she took during her degree. The reasons that G.P.A. is stupi rigorously, are the exact same reasons that G.P.A. seems stupid intuitively: college students get t choose their own courses, and the difficulty of getting an A relates to the difficulty of the cours you're in, so by choosing easier courses you can get better grades.

As a freshman and sophomore I didn't worry too much about G.P.A. because I assumed no-on (not employers, surely not grad schools) would be stupid enough to assign any meaning to a metr

which so blatantly failed to capture what it claimed to. But hey, apparently nobody is immune fro
from Old Man Endogeneity's wily tricks.

When people treat G.P.A. as if it actually captures something meaningful — for example, whe
they create a G.P.A. cutoff for a job posting or a grad school application — they are implicitl
assuming something like this:

X(f) + Y(b) = G.P.A. + error

f = effort
b = ability
X & Y are functions

The equation claims that G.P.A. is largely determined by effort and ability, plus or minus som
random variation. For an individual student, the error term here would sweep up such factors
whether she got woken up by a parrot at 4 a.m. the night before an exam, or whether her laptop died
the middle of writing her term paper. These causes of variation are not correlated with a student
ability or effort. If this model were correct then college G.P.A. would still not be a perfectly 'fai
measure, on an individual level, since some students would just be unlucky and get lower scores tha
their effort and ability merited. For all the students in the system put together, however, it would b
true that a high-G.P.A. student would with high probability possess better ability and effort than
low-G.P.A. student.

Once students are allowed to pick their own courses, however, G.P.A. becomes an implausibl
measure of effort and ability. While it may remain true that, for any given course you take, you
grades will correlate fairly well with the effort you put in and your ability at that subject, the fact th
you can choose your own courses introduces an opposing force: if you choose easier courses you ca
exert less effort and still attain better grades. As such, G.P.A. could correlate *inversely* with effort.

This, then, is the root of our endogeneity problem. Our error term (the difference between th
G.P.A. values predicted by our first model and the actual G.P.A. outcomes) is not at all random. I
fact, it is highly correlated with an omitted variable — difficulty of courses chosen. For any give
level of effort and ability, taking easier courses will result in a higher G.P.A. Note that even if ther
are brilliant, hard-working students who get 3.8s despite taking difficult courses, they would'v
(generally) done even better if they had taken easier courses.

Taking into account course selection, we can see that a more accurate equation for G.P.A. woul
look something like this:

X(f) + Y(b) + Z(c) = G.P.A. + error

f = effort
b = ability
c = easiness of courses selected
X, Y & Z are functions

This is still an over-simplification: there are a million other factors that determine G.P.A., fro
the all-important skill called "knowing-the-system" to the undoubtedly-influential "whether or n
student already has a job-offer by senior year." But our improved, three-input equation at least show
that a high G.P.A. can be caused not only by high ability and effort, which are admirable, but also b
purposefully taking easy courses, which is the-opposite-of-admirable. Someone who understands th

would be reticent to draw any kind of inference from a student's G.P.A. alone; she'd see that th signal given by G.P.A. is worse than useless.

Causality loops

A second kind of endogeneity problem occurs when cause and effect are connected by a kind o "causality loop." Endogeneity will be a problem whenever the output that you're trying to expla might also be the cause, rather than the effect, of the input that you use to explain it. When yo mother tells you she shouts at you because you never come to visit her, you might retort that you don like to visit because she's always shouting at you. "Mother," you tell her, "my visits are endogenou to how much you shout at me." Her implicit equation is:

$$\mathbf{A(1/v) = s + error}$$

v = number of times you visit,
s = number of times she shouts at you,
A = some function of 1/v.

In her model, the number of times she shouts at you is inversely related to the number of tim you visit: the more you visit, the less she shouts. But you point out that all her shouting causes you n to visit:

$$\mathbf{X(1/s) = v + error}$$

In your model, the number of times you visit is inversely related to the number of times she shou at you. So the amount you get shouted at is endogenous to how much you visit, and the amount yo visit is endogenous to how much you get shouted at. The upshot is that we can't draw any reliab inferences about how much you will visit (or be shouted at) from the models written above; we'd nee to create a better model that could handle the endogeneity problem properly.

People who switched

A similar problem plagues insurance advertising. Insurance firms regularly boast that "people wh switched their insurance to us saved an average of $102!" This tidbit tells us exactly nothing about th probability that we would save money by switching to their insurance. The sample of people wh switched insurance is massively endogenous; people who discover that they'd save a lot might switc providers, but people who discover that the new company would be more expensive for them a probably going to stay put. In a world where two insurance companies offered exactly the sam average prices, with a random assortment of people finding that they could save money by switchin from A to B but the same number able to save just as much by switching from B to A, *both* compani could honestly advertise that "people who switched their insurance to us saved an average of $X!" terms of endogeneity, the outcome to be explained (saving money) is actually a cause of the insuranc firms' explanation of that outcome (switching insurance providers), so the inference they unsubt imply (that you, or any other random person, will probably save money by switching to the insurance) is completely false.

In some ways we should be more impressed by a company that advertised that “people wh switched their insurance to us paid an average of $40 MORE.” This implies that people found th product so excellent that they were willing to pay more to get it, which at least makes it wor spending some time to find out more about it.

Bill and Mark are endogenous

Endogeneity is also the reason that you should never trust anyone who tells you something lik “Bill Gates and Mark Zuckerberg both dropped out of college and did very well for themselve therefore, we need to encourage our young-folks to be more adventurous, risk-takin entrepreneurial.” Bill and Mark were not just two randomly-selected Harvard undergrads — first, th fact that they dropped out of college was endogenously determined by their personalities, which als affected their probability of success; second, the fact that they are known about by the person givi the example is itself endogenous to their later successes.

For Bill and Mark, part of the endogeneity was from *personality*. What kind of person drops out Harvard aged 20? Well, suffice to say that it’s not an ordinary Harvard undergrad, given th graduation rate. In a world where dropping out of college is a big taboo, and where going to Harvard kind-of-a-big-deal, the sort of person who drops out of Harvard is not an ordinary sort of person at a (or an ordinary sort of Harvard undergrad, which is far from the same thing). Bill and Mark probab both had some extraordinary personality traits — boldness, courage, hacking ability, self-belief - which helped them greatly in their business careers, *and* which made them willing to drop out Harvard.

In Mark’s case there was an additional source of drop-out-decision endogeneity: he already had pretty successful company, with Silicon Valley offices and Peter Thiel’s investment, before h officially dropped out of school. The omitted variable — whether or not your company is alread mind-blowingly successful — is surely correlated to the probability that you’ll decide to drop out college, and your decision to drop out of college will then seem to have more of a positive influen on your successes than it really should. People who just note that Mark was a risk-taking colleg dropout, but fail to note that he was already massively successful before he dropped out, will ove estimate the benefit of dropping out for those *without* an already-successful company.*

The first type of endogeneity problem, discussed above, occurs at the level of deciding to drop ou But there is a second endogeneity problem that occurs at the level of people-discussing-colleg dropouts. Most college dropouts are not well-known enough that strangers would discuss them wh talking about college dropouts. Imagine if your Cousin Frank abandoned Harvard and went on to liv in his girlfriend’s garage for the rest of his life, running a failed internet startup. *You* would kno about Cousin Frank’s misadventures, and the rest of Frank’s family would, and some of his high school friends. Let’s say, generously, that 1,000 people would be aware of Frank’s dropout status an his unsuccessful existence (maybe through reading his Facebook updates). By contrast, millions ar millions of people know about Mark Zuckerberg’s story, how he dropped out of Harvard and became billionaire: 20 million officially saw the Zuckerberg biopic *The Social Network*, and 500 million hav accounts with his website. Taking the low-ball 20 million figure, for every one Mark Zuckerberg the could be 20,000 Cousin Franks, and (if none of the Cousin Franks had any mutual friends, which is stupid assumption but let it slide for now) then the 20 million people who knew of one Ma Zuckerberg and one Cousin Frank might assume that college droupouts had a 50/50 chance becoming billionaires. Not bad odds, I’d say. But of course this calculation is junk: the ‘actual’ odd (in our artificial example) would be 1 / 20,001. Not so hot.

Consultants

If endogeneity can fool ordinary people, it can also apparently fool huge multinational companie
A leading American management-consulting firm has a prominent graph on its website showing ho
their clients' share-price gains since 1980 have outperformed a stock-market index by a factor of fou
This looks pretty impressive, unless you think about it. There's two endogeneity effects here. First, t
kind of client who tries to hire a consultant is not representative of companies as a whole — perha
the companies who hire consultants have more dynamic leadership to begin with; perhaps the future
looking bright for them, and they feel that they have cash to spare; or perhaps they were doi
unusually poorly, needed restructuring, but were about to bottom-out anyway. Second, the clients th
a consulting firm accepts are not a random sample either: a consulting firm may only take on clien
that it thinks it can help, or perhaps even that it thinks will have share-price gains in future years (
opposed to one that will still do badly, but perhaps less-badly, with the consultants' help). If th
consulting firm had randomised which clients it accepted and which it didn't, the difference in sha
price gains between the accepted and rejected firms would provide a meaningful gauge of t
consultants' contribution. The statistics presented on the website, however, tells us more about wh
hires consultants than about how much consultants help companies.

Social science and natural science

One of the great problems of social science is that nothing ever sits still. Modern statistics got i
big break with 1920s agricultural experiments: you take two nearby rows of corn and treat them wi
two different fertilizers, and at the end of the season you see which field grew better. The early yea
in the development of statistics were largely about figuring out the quality of inferences you can mal
from different kinds of data: how do you tell if the different yields were caused by a better fertiliser
just a fluke? How many ears of corn do you need to check before you can confidently declare t
characteristics of the entire field? These are difficult questions, and it took some obscenely sma
people to provide rigorous answers. But one comfort we had was that the direction of causality w
completely clear. The independent variable (which fertiliser you used) would affect the depende
variable (how much corn grows), but there was absolutely no way that the dependent variable (co
growth) could influence the independent variable (which of the fertilisers you had already used on it)

With humans involved, nothing is ever so simple. Suppose you're interested in primary schoolin
in developing countries, and want to see the effects that donating textbooks has on the student
educational outcomes. You find a village with two identical schools in it and give textbooks only t
the students in School A. So far, so good: you have an independent variable (whether a student got
textbook) and a dependent variable (the students' educational outcomes at the end of the study). B
unlike corn-stalks, children can *move*. If they (or their parents) start to realise that one school
getting better resources than the other, some of the more ambitious students might try to move acros
The end of the study arrives and the students in School A achieved better academic outcomes. But di
the textbooks cause the better outcomes, or did the more ambitious students cause themselves to g
given textbooks? How much of the gains are due to each cause? It's hard to tell. The supposedly
dependent variable has started influencing the supposedly-independent ones. Cause and effect hav
become difficult to disentangle.

In fact, endogeneity plagues all kinds of social science research, and many of the social scien

questions that make it into the news. If you see a piece of social science reported in the media, one your first questions should probably be "is there an endogeneity problem here?" This is n necessarily the first question you should ask about social science research *in general*, but there seem to be a kind of endogeneity problem with the kind of social science that gets reported in the medi namely that the kind of papers that make it into the news are unusually likely to suffer fro endogeneity problems. To a stats nerd, this passes as pretty funny.

Not all that correlates is caused

If there's one statistical maxim which is widely known among non stats-nerds it's "correlatio does not imply causation." *Correlation* means, roughly, that two variables are inter-dependent: if on goes up, the other goes up with it. (If two variables are *inversely correlated* then as one goes up, th other goes down). *Causation* means, well, that one *caused* the other to happen. But two variables ca be correlated without any causal relationship between them: it could just be coincidence, or it could b that they are both caused by an invisible third factor. Just because two things vary together does n mean one caused the other.

A classic example of "correlation does not imply causation" is the famous story that ice-crea sales over the course of a year tend to correlate with the number of drownings. Does this mean tha say, eating ice-cream causes significant groups of children to go sugar-crazy and fall in a lake? C even more bizarrely, that while people are drowning they suddenly consume a lot of ice-cream? Wel unsurprisingly, no. Ice-cream sales tend to go up in summer, a time when people also spend more tim swimming outdoors, so rising ice-cream sales and increased drownings are both caused by warm weather but aren't actually related directly.

If you think this is silly and that no-one would make such a basic causation-and-correlatio mistake: needless to say, think again. Public Health experts in the 1940s noticed a correlation betwee polio cases and ice-cream consumption; they recommended cutting out ice-cream to protect again the disease. It later turned out that, you guessed it, polio outbreaks were more common in summe and ice-cream eating was more common in summer, and polio and ice-cream had nothing to do wi each other.*

Another great example is that the average price of rum is strongly correlated with the averag wages of kindergarten-teachers. Do rising childcare wages give more spare cash to exasperate teachers, bidding up rum prices? Or do higher rum prices cause people to drink less, therefore wo harder, therefore spend more money on their childrens' kindergartens? Neither, obviously: the price o basically-everything rises over time due to inflation, and this extraneous factor affects kindergarte teachers' wages and rum prices and the cost of houses and MPs' salaries and almost-everything els You can show correlation between any two price-rises and, if you're a devious sort of person, ignor the fact that both increases are caused by an invisible third factor, the general fact that prices tend rise over time. In fact, lots of trends tend to increase over time, and if you want to get really silly yo can start to point out correlations between *any two* such variables, ignoring the invisible third fact (that many things increase over time, in general).

A well-known newspaper once wrote about a "striking correlation" between the number of mil driven per licensed American driver and U.S. obesity rates six years later— the hypothesis was th when Americans drove more they exercised less, and (following a time-lag to allow for the change affect physiques) they got fat. The newspaper, although straining to note that "these predictions com with a strong caveat: correlation does not equal causation," nonetheless seemed very impressed at th "near-perfect correlation." A famous economist, Justin Wolfers, went one better and showed an even

better correlation between American obesity rates and his age. As Wolfers writes: “I’m not arguin that my aging is causing higher obesity. Rather, when you see a variable that follows a simple tren almost any other trending variable will fit it: miles driven, my age, the Canadian population, tot deaths, food prices, cumulative rainfall, whatever.”*

Correlation, causation and endogeneity

At this point, the astute reader might wonder to herself: “couldn’t all the examples above l expressed as endogeneity problems?” At which the proud author would sob loudly into his laptop, an tell everyone who would listen about how quickly Astute Reader had grown up, and wasn’t it onl yesterday that she was browsing the introduction? Yes, indeed, correlation-is-not-causation is specific example of an endogeneity problem. Take ice-cream and drownings. Our original (incorrec formula was:

A(ice-cream sales) = drownings + error

Our omitted variable was “outdoor temperature,” which we’ve lumped into our error term b which is in fact correlated with both ice-cream sales and drownings. The correct formulas would l (something like):

Y(temp.) = ice-cream sales + error
and
Z(temp.) = drownings + error

with any apparent connection between ice-cream sales and drownings actually driven by th common independent variable “outdoor temperature.” Correlation is not causation: both outcomes a endogenous to an omitted variable that has been wrongly lumped into the error term.

CHAPTER THREE: BAYES

Bayes matters

You come home early one day to find that your girlfriend, who told you she was going hom because she wasn't feeling well, actually snuck out to dinner with her ex. (I hate to be the one to te you this, but really, you had to know). What is the probability that she's cheating on you? Mo importantly: how should you *calculate* the probability that she's cheating on you? What factors do yo need to take into account? You won't, and shouldn't, assign numbers here; they'd be thoroughl arbitrary ("Honey, there seems to be a 23.97% probability that you're cheating on me. Plus or minu 0.01%"). But you're going to make an implicit probabilistic assessment anyway ("She's definitel cheating on me;" "I think she might be cheating on me;" "I'm only slightly worried that she cheating on me;" etc. etc.), so you might as well make the assessment good.

The question here is something called a *conditional probability*, that is, the probability of X give Y — here, the probability that she's cheating *given* that she snuck off to dine with her ex. So long two things are in some way related, the conditional probability for X given Y will not be the same a the simple probability of X all on its own. For instance, the simple probability that a randomly-chose burrito is going to be tasty might be 0.4. The conditional probability that it's going to be tasty *give* that your friend recommended this particular restaurant might be 0.8 — her recommendation is a no random indicator of burrito quality. And the conditional probability that the burrito will be tasty *give* that your friend recommended the place, *and* that you've previously eaten quesadillas here and the were excellent, *and* that it won some prestigious burrito-making award, might be 0.99. Statisticia use the symbol P(X|Y) to represent the probability of X happening given that Y did; the vertical ba represents the phrase "given that."

Now to return to our girlfriend-problem. We have a hypothesis about the world (the hypothes that she might be cheating on us) and a new piece of evidence (she snuck off to dinner without telli us). The Reverend Thomas Bayes was an 18th Century British clergyman who figured out exactly ho to deal with these kinds of problems, where we want to know the probability that our hypothesis (th is, our causal explanation for *why* something happened) is correct given a new piece of evidence (th is, something that has happened). Bayes' insight was that the conditional probabili P(Hypothesis|Evidence) depends on four different things— well, they're kind of connected so there not technically four of them, but let's call it four different things for now.

First, P(Hypothesis|Evidence) depends on the probability P(Evidence|Hypothesis): if a piece evidence was extremely probable under a given hypothesis, the existence of the evidence makes more probable that the hypothesis was correct. This makes sense: if our hypothesis strongly predicte that a particular piece of evidence would show up, and that piece of evidence does indeed show u then it is a little more probable that our hypothesis was correct (because if our hypothesis ha predicted a particular piece of evidence, and that evidence had *not* shown up, then we'd be forced conclude that our hypothesis was probably wrong).

Second, it depends on the probability that the hypothesis was correct *before* we saw the ne evidence; statisticians call this the "prior probability." New evidence can be used to help "update" ou beliefs — that is, after seeing a new piece of evidence, our previous hypothesis can become more

less probable. But we must never forget the prior probability that we started with. If the star striker o your favourite soccer team gets injured—new evidence—then it becomes less probable than it ha been that your team will win the next match. However, the *absolute* probability that they will win th next match is still very dependent on how good they are, and how good the other team is, and ho probable their win was before the injury came along. The new evidence (the striker's injury) can on cause to update our prior beliefs; it doesn't act in a vacuum, and couldn't possibly tell us ho probable a win is in the absence of prior beliefs.

The third and fourth factors for the Bayesian analysis are mirrors of the first two, but with regar to the *alternative hypotheses* that compete with our own one to explain the world around us. Th alternative hypotheses are all the other possible ways to explain what happened; all the things *apa from* our hypothesis that could have caused the evidence to arise. For example, perhaps you go to neighbours house and the family dog jumps up and licks you. "The dog is licking me!," you exclaim "this new evidence supports my hypothesis that the dog really likes me." But as a good statistician yo must also consider alternative hypotheses: for example, that the dog licks everyone, or that yo aftershave smells like meat. These alternatives, and many others, could also explain the eviden you're seeing.

The third factor, then, is the probability of the new evidence given the alternative hypotheses, an the fourth factor is the prior probabilities of each of the alternative hypotheses. Essentially, the sam arguments as above apply in reverse: the more probable the evidence under alternative hypothese and the more probable those hypotheses were to begin with, the more probable it is that the ne evidence has appeared due to these alternative causes, and so the *less* probable it is that our origin hypothesis is correct. Conversely, the new evidence argues much more strongly for our initi hypothesis if the piece of evidence is highly *improbable* under the reasonable alternative hypothese or if those alternative hypotheses were initially very improbable.

Don't worry if that all seemed confusing, or if you're not sure you absorbed what you just rea unfortunately Bayes Theorem' is just one of those things you have to chew on for a while before makes sense. Let's look at a more concrete example. Suppose that a note shows up on my des purporting to reveal a secret crush from the most beautiful girl in the class. It is indeed somewha probable that the girl would send such a message if she had a crush on me: P(message|crush) moderately-high. Depending on how attractive I am, perhaps P(crush) — the probability that sh really does have a crush on me — is also non-negligible. Unfortunately, there are still the alternativ hypotheses to consider: for example, that my friends are playing a prank. The probability that I wou receive such a message if my friends were playing a prank — P(message|prank) — is fairly high, an the probability that my friends would play a prank on me, P(prank), is extremely high. This mean that, overall, the existence of the message becomes weak evidence for the hypothesis that the girl ha a crush on me; the alternative hypotheses explain the evidence just as well, and were more probab before the evidence showed up.

Formula for Bayes

Reverend Bayes was in fact the source for more than this little theorem — his work was th cornerstone for an entire statistical philosophy commonly known as (this is going to shock yo Bayesian Statistics. There is an enormous amount of controversy in the statistics world about wheth Bayesianism is really the correct way to think about probabilities, as opposed to another school thought called Frequentism, and if you ever find yourself at a statistics conference you must promi (promise!) not to mention that this book flagrantly side-stepped the Bayesian-Frequentist debate -

this is one of those questions that inspires an enormous amount of passion and argument from th people who really care about it. For our purposes, though, the philosophical debate is not critica Bayes' Theorem is a great tool for thinking about probabilities, regardless of deeper questions abo the true nature of probabilities in our universe. Bayes' Theorem can be expressed as a simple formul take a deep breath, it's honestly (honestly!) not as bad as it looks:

$$\mathbf{P(Hypothesis_1|Evidence)} = \frac{\mathbf{P(Evidence|Hypothesis_1)P(Hypothesis_1)}}{\mathbf{P(Evidence)}}$$

P(X) = the probability of X occurring,
*P(X|Y) = the probability of X occurring **given** that Y occurred.*

What does that mean? Well, let's return to our example with the note from the beautiful girl liked that example). Our $Hypothesis_1$ was that I received the note because the girl had a crush on m and the probability we're trying to assess is $P(H_1|E)$: the probability that she really does have a crus on me given that I received the note. $P(E|H_1)$ was the probability that she would send the note if sh had the crush; $P(H_1)$ was the probability that she had a crush on me before we knew the note existe P(E) is the total probability of seeing the evidence under all possible hypotheses; we get it by addi together $P(E|H_1)P(H_1)$ and $P(E|H_2)P(H_2)$ and $P(E|H_3)P(H_3)$, etc. etc., until we have gone through a the reasonable alternative hypotheses: the total probability of seeing the evidence is a function of ho probable the evidence was in each possible state of the world, weighted by how probable that state the world was to begin with, so we arrive at P(E) by summing P(E|H)P(H) for all possible H's. It crucial that we include our own hypothesis, H_1, in the calculation of P(E) — after all, our hypothes is one of the relevant states of the world under which the new evidence might occur.

Assimilating evidence

Now that we know the formula, we can start to explore its consequences. We know that there ar four elements to consider when thinking about a conditional probability: the probability of E give H_1; the prior probability for H_1; the probability of E given H_{others}; and the prior probabilities f H_{others}. As previously discussed, if $P(E|H_1)$ rises then $P(H_1|E)$ also rises; this makes sense, because th more probable the new evidence under our hypothesis, the more that seeing the new evidence mak our hypothesis more probable itself. If the prior probability $P(H_1)$ rises then the posterior probabili $P(H_1|E)$ also rises; this makes sense, because the higher the hypothesis' probability was to begin wit the higher it will be after any given piece of new evidence is introduced. Conversely, if $P(E|H_{other}$ rises then the posterior probability $P(H_1|E)$ *falls*; this makes sense, because the more probable the ne evidence is under the alternative hypotheses, the less the existence of the new evidence makes u certain that our $Hypothesis_1$ was correct. Finally, if the posterior probability $P(H_{others})$ rises then th posterior probability $P(H_1|E)$ falls; this makes sense, because the more probable the alternati hypotheses were to begin with, the more probable they will be after our new evidence arrives, and s the less support the new evidence provides for our own original hypothesis — the new evidence mig be proof, instead, that one of the *other* hypotheses was correct.