4
Simpson's paradox and the rule of total probability

Simpson's paradox

CONSIDER the following data on obstetricians delivering babies at a hospital in England. The table below shows the complication rates for both junior and senior doctors on the delivery ward, grouped by delivery type:

	Easier deliveries	Harder deliveries	Overall
Senior doctors	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctors	0.067 (3169)	0.155 (206)	0.072 (3375)

The numbers in parentheses are the total deliveries of each type.

This table exhibits an aggregation paradox, also called *Simpson's paradox.* No matter what kind of delivery you have, whether easy or hard, you'd prefer to have a senior doctor. They have lower complication rates than junior doctors in both cases. Yet counterintuitively, the senior doctors have a higher overall complication rate: 7.6% versus 7.2%. Why? Because of a lurking variable: most of the deliveries performed by junior doctors are easier cases, where complication rates are lower overall. The senior doctors, meanwhile, work a much higher fraction of the harder cases. Their overall complication rate reflects this burden.

Here's another example. Jacoby Ellsbury and Mike Lowell were two baseball players for the Boston Red Sox during the 2007 and 2008 seasons. The table below shows their batting averages for those two seasons, with their number of at-bats in parentheses. We see that Ellsbury had a higher batting average when he was a rookie, in 2007; a higher batting average a year later, in 2008; but a lower batting average overal!

Again we have an aggregation paradox, and again it is resolved by pointing to a lurking variable: in 2007, when both players had

	2007	2008	Overall
Lowell	.324 (589)	.274 (419)	.304 (1008)
Ellsbury	.353 (116)	.280 (554)	.293 (670)

higher averages, Ellsbury had many fewer at-bats than Lowell.

It turns out the math of these aggregation paradoxes can be understood a lot more deeply in terms of something called the *rule of total probability*, or the *mixture rule*. This rule sounds impressive, and we've written out the full math version below (see Equation 4.2 on page 37). But the intuition of this rule is actually quite simple. It says: the probability of any event is the sum of the probabilities for all the different ways in which the event can happen. In that sense, the law of total probability is really just Kolmogorov's third rule in disguise. The distinct ways in which some event A can happen are mutually exclusive. Therefore we just sum all their probabilities together to get $P(A)$.

Let's return to the example on obstetric complication rates on junior doctors at a hospital in England. In the table, there are two ways of having a complication: with an easy case, or with a hard case. Therefore, the total probability is the sum of two joint probabilities:

$$P(\text{complication}) = P(\text{easy and complication}) + P(\text{hard and complication})\,.$$

If we now apply the rule for conditional probabilities (Equation 2.1) to each of the two joint probabilities on the right-hand side of this equation, we have this:

$$P(\text{complication}) = P(\text{easy}) \cdot P(\text{complication} \mid \text{easy}) + P(\text{hard}) \cdot P(\text{complication} \mid \text{hard})$$

Thus the rule of total probability says that overall probability is a weighted average—a mixture—of the two conditional probabilities. For senior doctors we get

$$P(\text{complication}) = \frac{213}{315} \cdot 0.052 + \frac{102}{315} \cdot 0.127 = 0.076\,.$$

And for junior doctors, we get

$$P(\text{complication}) = \frac{3169}{3375} \cdot 0.067 + \frac{206}{3375} \cdot 0.155 = 0.072\,.$$

This is a lower *overall* probabiity of a complication, despite the fact the junior doctors have higher conditional probabilities of a complication in all scenarios.

So which probabilities should we report: the conditional probabilities, or the overall (total) probabilities? There's no one right answer; it depends on your conditioning variable, and your goals. In the obstetric data, the overall complication rates are clearly misleading. The distinction between easier and harder cases matters a lot. Senior doctors work harder cases, on average, and therefore have higher overall complication rates. But what matters to the patient, and to anyone who assesses the doctors' performance, are the *conditional* rates. You have to account for the lurking variable.

The baseball data is different. Here the *conditional* probabilities for 2007 and 2008 are probably misleading. The distinction between 2007 and 2008 is nothing more than an arbitrary cutoff on the calendar. It's barely relevant from the standpoint of assessing baseball skill, and it needlessly splits one big sample of each player's history into two smaller, more variable samples. So in this case we'd probably go with the overall averages if we wanted to say which player was performing better.

A formal statement of the rule of total probability

Suppose that events $B_1, B_2, \ldots, B_N$ constitute an exhaustive partition of all possibilities in some situation. That is, the events themselves are mutually exclusive, but one of them must happen. This can be expressed mathematically as

$$P(B_i, B_j) = 0 \text{ for any } i \neq j, \quad \text{and} \quad \sum_{i=1}^{N} P(B_i) = 1\,. \tag{4.1}$$

Now consider any event A. If Equation 4.1 holds, then

$$P(A) = \sum_{i=1}^{N} P(A, B_i) = \sum_{i=1}^{N} P(B_i) \cdot P(A \mid B_i)\,. \tag{4.2}$$

Equation 4.2 is what is usually called the rule of total probability.

Surveys and the rule of total probability

One of the least surprising headlines of 2010 must surely have been the following, from the ABC News website:

> **Teens not always honest about drug use.**[1]

[1] Kim Carollo, ABC News, Oct. 25, 2010. Link here.

In other news, dog bites man.

To be fair, the story itself was a bit more surprising than the headline. Yes, it's hardly news that teenagers would lie to their parents, teachers, coaches, and priests about drug use. But the ABC News story was actually reporting on a study showing that teenagers also lie to researchers who conduct anonymous surveys about drug use—even when those teenagers know that their answers will be verified using a drug test.

Here's the gist of the study. Virginia Delaney-Black and her colleagues at Wayne State University, in Detroit, gave an anonymous survey to 432 teenagers, asking whether they had used various illegal drugs.[2] Of these 432 teens, 211 of them also agreed to give a hair sample. Therefore, for these 211 respondents, the researchers could compare people's answers with an actual drug test.

The two sets of results were strikingly different. For example, of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine. However, when the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

And it wasn't just the teens who lied. The survey researchers also asked the *parents* of the teens whether they themselves had used cocaine. Only 6.1% said yes, but 28.3% of the hair samples came back positive.

Let's emphasize again that we're talking about a group of people who were guaranteed anonymity, who wouldn't be arrested or fired for saying yes, and who willingly agreed to provide a hair sample that they knew could be used to verify their survey answers. Yet a big fraction lied about their drug use anyway.

Surveys and lies

Drug abuse—whether it's crack cocaine in Detroit, or bathtub speed in rural Nebraska—is a huge social problem. It fills our jails, drains public finances, and perpetuates a trans-generational cycle of poverty. Getting good data on this problem is important. As it stands, pediatricians, schools, and governments all rely on self-reported measures of drug use to guide their thinking on this issue. Yet distressingly, the proportion of self-reported cocaine use in the Detroit study, 0.7%, was broadly similar to the findings in large, highly regarded surveys—for example, the federally funded National Survey on Drug Use and Health. The work of Dr. Delaney-Black and her colleagues would seem to imply that all

[2] V. Delaney–Black et. al. "Just Say 'I Don't': Lack of Concordance Between Teen Report and Biological Measures of Drug Use." *Pediatrics* 165:5, pp. 887-93 (2010).

of these self-reported figures might be way off the mark.

Moreover, theirs hasn't been the only study to uncover evidence that surveys cannot necessarily be taken at face value. Here are some other things that, according to research *on* surveys, people lie about *in* surveys.

- Churchgoers overstate the amount of money they give when the hat gets passed around during the service.
- Gang members embellish the number of violent encounters they have been in.
- Men exaggerate their salary, among other things.
- Ravers will "confess" to having gotten high on drugs that do not actually exist.

How to ask an embarrassing question: probability as an invisibility cloak

But there's actually some good news to be found here. It's this: when people lie in surveys, they tend to do so for predictable reasons (to impress someone or avoid embarrassment), and in predictable ways (higher salary, fewer warts). This opens the door for survey designers to use a bit of probability, and a bit of psychology, to get at the truth—even in a world of liars.

Let's go back to the example of drug-use surveys so that we can see this idea play out. Suppose that you want to learn about the prevalence of drug use among college students. You decide to conduct a survey at a large state university to find out how many of the students there have smoked marijuana in the last year. But as you now appreciate, if you ask people direct questions about drugs, you can't always trust their answers.

Here's a cute trick for alleviating this problem, in a way that uses probability theory to mitigate someone's psychological incentive to lie. Suppose that, instead of asking people point-blank about marijuana, you give them these instructions.

1. Flip a coin. Look at the result, but keep it private.
2. If the coin comes up heads, please use the space provided to write an answer to question Q1: "Is the last digit of your Social Security number odd?"
3. If the coin comes up tails, please use the space provided to write an answer to question Q2: "Have you smoked marijuana in the last year?"

The key fact here is that only the respondent knows which question he or she is answering. This gives people plausible deniability. Someone answering "yes" might have easily flipped heads and answered the first, innocuous question rather than the second, embarrassing one, and the designer of the survey would never know the difference. This reduces the incentive to lie.

Moreover, despite the partial invisibility cloak we've provided to the marijuana users in our sample, we can still use the results of the survey to answer the question we care about: what fraction of students have used marijuana in the past year? We'll use the following notation:

- Let Y be the event "a randomly chosen student answers yes."
- Let Q_1 be the event "the student provided an answer to question 1, about their Social Security number."
- Let Q_2 be the event "the student provided an answer to question 2, about their marijuana use."

From the survey, we have an estimate of $P(Y)$, which is the overall fraction of survey respondents providing a "yes" answer. We really want to know $P(Y \mid Q_2)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the marijuana question. The problem is that we don't know which students were answering the marijuana question.

To understand the rule of total probability, let's return to our hypothetical survey in which we want to know the answer to the question: what fraction of students have used marijuana in the past year? Then we have each survey respondent privately flip a coin to determine whether they answer an innocurous question (Q1) or the question about marijuana use (Q2). We used the following notation:

- Let Y be the event "a randomly chosen student answers yes."
- Let Q_1 be the event "the student provided an answer to question 1, about their Social Security number."
- Let Q_2 be the event "the student provided an answer to question 2, about their marijuana use."

To solve this problem, we'll use rule of total probability. In the case of our drug-use survey, this means that

$$P(Y) = P(Y, Q_1) + P(Y, Q_2)\,. \tag{4.3}$$

In words, this equation says that there are two ways to get a yes answer: from someone answering the social-security-number question, and from someone answering the drugs question. The total number of yes answers will be the sum of the yes answers from both types in this mixture.

Now let's re-write Equation 4.3 slightly, by applying the rule for conditional probabilities to each of the two joint probabilities on the right-hand side of this equation:

$$P(Y) = P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2) \, . \qquad (4.4)$$

This equation now says that the overall probability $P(Y)$ is a weighted average of two conditional probabilities:

- $P(Y \mid Q_1)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the social-security-number question.

- $P(Y \mid Q_2)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the marijuana question.

The weights in this average are the probabilities for each question: $P(Q_1)$ and $P(Q_2)$, respectively.

Now we're ready to use Equation 4.4 to calculate the probability that we care about: $P(Y \mid Q_2)$. We know that $P(Q_1) = P(Q_2) = 0.5$, since a coin flip was used to determine whether Q_1 or Q_2 was answered. Moreover, we also know that $P(Y \mid Q_1) = 0.5$, since it is equally likely that someone's Social Security number will end in an even or odd digit.[3]

We can use this information to simplify the equation above:

$$P(Y) = 0.5 \cdot 0.5 + 0.5 \cdot P(Y \mid Q_2) \, ,$$

or equivalently,

$$P(Y \mid Q_2) = 2 \cdot \{P(Y) - 0.25\} \, .$$

Suppose, for example, that 35% of survey respondents answer yes, so that $P(Y) = 0.35$. This implies that

$$P(Y \mid Q_2) = 2 \cdot (0.35 - 0.25) = 0.2 \, .$$

We would therefore estimate that about 20% of students have smoked marijuana in the last year.

[3] This survey design relies upon the fact that the survey designer doesn't know anyone's Social Security number. If you were running this survey in a large company, where people's SSNs were actually on file, you'd need to come up with some other innocuous question whose answer was unknown to the employer, but for which $P(Y \mid Q_1)$ was known.