Warm-up

SDS 323
James Scott (UT-Austin)

A bit of probability and a bit of R to get the semester started.

Outline

1. Reminder of some probability basics
2. Conditional probability
3. Conditional probabilities from data
4. Rule of total probability
5. Independence
6. Bayes' Rule

Probability basics

If A denotes some event, then $P(A)$ is the probability that this event occurs:

- P(coin lands heads) = 0.5
- P(rainy day in Ireland) = 0.85
- P(cold day in Hell) = 0.0000001

And so on.

Probability basics

Some probabilities are estimated from direct experience over the long run:

- P(newborn baby is a boy) = $\frac{106}{206}$
- P(death due to car accident over a one year period) = $\frac{12}{100{,}000}$
- P(living forever) = 0

Probability basics

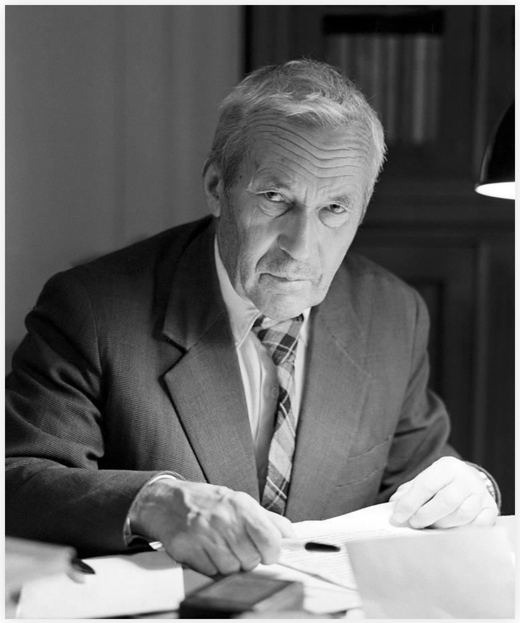
Some probabilities are estimated from direct experience over the long run:

- P(newborn baby is a boy) = $\frac{106}{206}$
- P(death due to car accident over a one year period) = $\frac{12}{100{,}000}$
- P(living forever) = 0

Others are synthesized from our best judgments about unique events:

- P(Apple stock goes up after next earnings call) = 0.54
- P(Djokovic wins next US Open) = 0.4 (6 to 4 odds)
- etc.

Kolmogorov's axioms (baby version)



“Obey my rules, filthy capitalists.”

Consider an uncertain outcome with where Ω is the set of all possible outcomes. “Probability” $P(\cdot)$ is a set function that maps Ω to the real numbers, such that:

1. Non-negativity: For any event $A \subset \Omega$, $P(A) \geq 0$.
2. Normalization: $P(\Omega) = 1$ and $P(\emptyset) = 0$.
3. Finite additivity: If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$.

Conditional probability

A conditional probability is the chance that one thing happens, given that some other thing has already happened.

A great example is a weather forecast: if you look outside this morning and see gathering clouds, you might assume that rain is likely and carry an umbrella.

We express this judgment as a conditional probability: e.g. "the conditional probability of rain this afternoon, given clouds this morning, is 60%."

Conditional probability

In stats we write this a bit more compactly:

- P(rain this afternoon | clouds this morning) = 0.6
- That vertical bar means “given” or “conditional upon.”
- The thing on the left of the bar is the event we're interested in.
- The thing on the right of the bar is our knowledge, also called the “conditioning event” or “conditioning variable”: what we believe or assume to be true.

$P(A \mid B)$: “the probability of A, given that B occurs.”

Conditional probability

Conditional probabilities are how we express judgments in a way that reflects our partial knowledge.

- You just gave *Sherlock* a high rating. What's the conditional probability that you will like *The Imitation Game* or *Tinker Tailor Soldier Spy*?
- You just bought organic dog food on Amazon. What's the conditional probability that you will also buy a GPS-enabled dog collar?
- You follow Lionel Messi (@leomessi) on Instagram. What's the conditional probability that you will respond to a suggestion to follow Cristiano Ronaldo (@cristiano) or Gareth Bale (@garethbale11)?

Conditional probability

A really important fact is that conditional probabilities are *not symmetric*:

$$P(A \mid B) \neq P(B \mid A)$$

As a quick counter-example, let the events A and B be as follows:

- A: “you can dribble a basketball”
- B: “you play in the NBA”

Conditional probability

- A: “you can dribble a basketball”
- B: “you play in the NBA”

Clearly $P(A \mid B) = 1$: every NBA player can dribble a basketball.

Conditional probability

- A: “you can dribble a basketball”
- B: “you play in the NBA”

But $P(B \mid A)$ is nearly zero!

The multiplication rule

We've met Kolmogorov's three axioms, but there's one final axiom for conditional probability, often called the *multiplication rule*.

Let $P(A, B) = P(A \cap B)$ be the *joint probability* that both A and B happen. Then:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} .$$

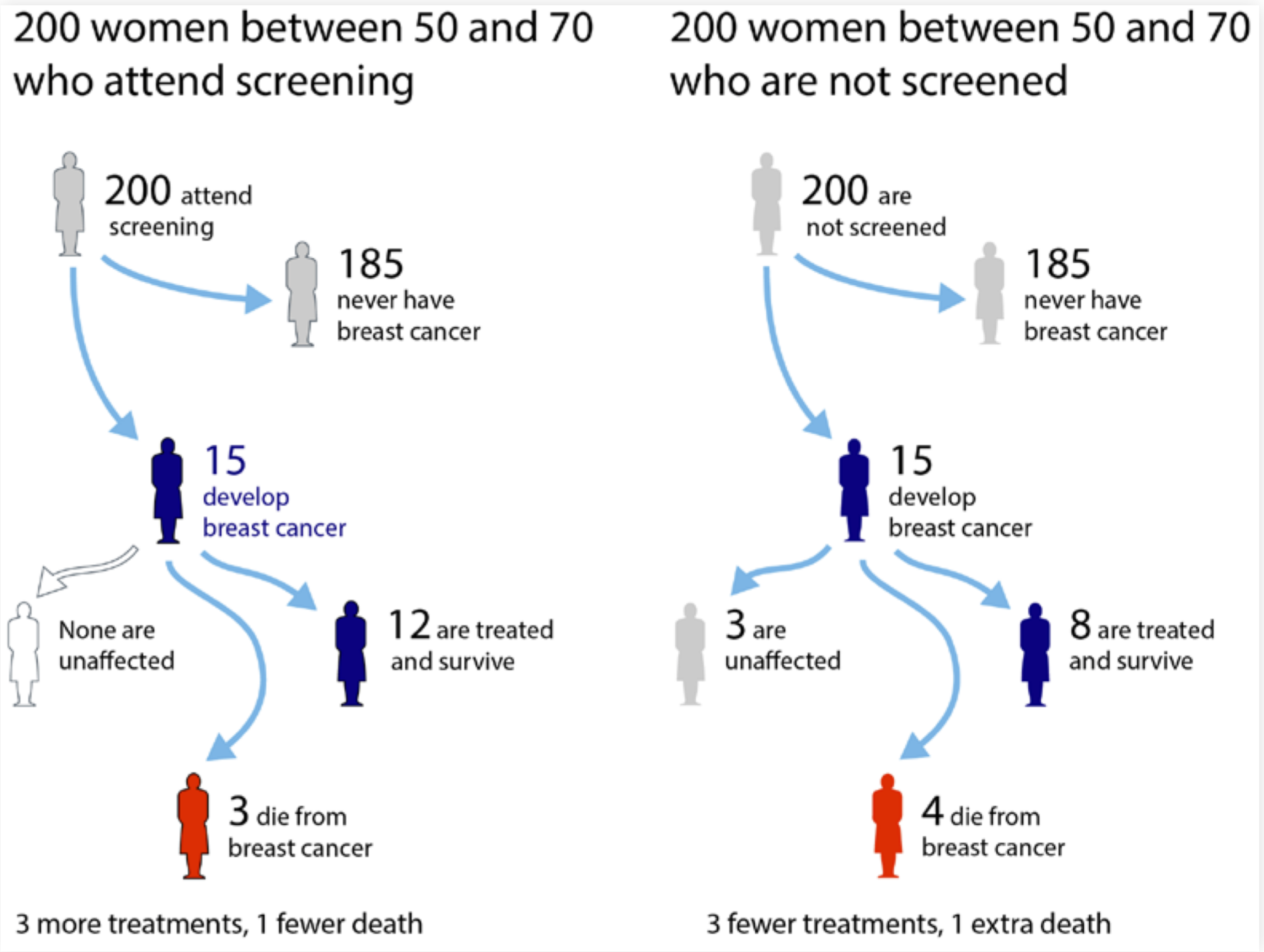
Or equivalently:

$$P(A, B) = P(A \mid B) \cdot P(B) .$$

This is an axiom: it cannot be proven from Kolmogorov's rules.

Conditional probability

Let's see why this axiom makes sense. (Figure courtesy David Speigelhalter and Jenny Gage.)



Conditional probability

Suppose a woman goes for regular screening (left branch). What is P(survive | cancer)?

- Among screened women, 15 will get cancer, on average.
- Of those 15, 12 are treated and survive, on average.
- Thus inuitively, we should have P(survive | cancer) = 12/15 = 0.8

We get the same answer using the rule for conditional probabilities:

$$P(S|C) = \frac{P(S,C)}{P(C)} = \frac{12/200}{15/200}$$
$$= \frac{12}{15}$$

Conditional probabilities from data

Suppose that you're designing the movie-recommendation algorithm for Netflix, and you have access to the entire Netflix database, showing which customers have liked which films.

Your goal is to leverage this vast data resource to make automated, personalized movie recommendations.

You decide to start with an easy case: assessing how probable it is that a user will like the film *Saving Private Ryan* (event A), given that the same user has liked the HBO series *Band of Brothers* (event B).

This is almost certainly a good bet!

Conditional probabilities from data

But keep in mind that you want to be able to do this kind of thing automatically.

Key insight: *frame the problem in terms of conditional probability.*

- Suppose we learn that you liked film B, but haven't yet seen film A.
- What is $P(\text{likes A} \mid \text{likes B})$? *Conditional probabilities hold the key to understanding individualized preferences.*
- So how can we learn $P(\text{likes A} \mid \text{likes B})$?

Conditional probabilities from data

Solution: go to the data! Suppose your database on 5 million subscribers like Linda reveals the following pattern:

	Liked *Band of Brothers*	Didn't like
Liked *Saving Private Ryan*	2.8 million	0.3 million
Didn't like	0.7 million	1.2 million

Then

$$
\begin{aligned}
P(\text{liked Saving Private Ryan} \mid \text{liked Band of Brothers}) &= \frac{2.8 \text{ million}}{3.5 \text{ million}} \\
&= 0.8\,.
\end{aligned}
$$

Result: a good recommendation with no human in the loop.

Conditional probabilities from data

Many companies do the same:

- Amazon for products
- New York Times for new stories
- Google for web pages
- etc

The digital economy runs on conditional probability.

Conditional probabilities from data

Let's try an example ourselves in `predimed_intro.R` from the class website.

We'll use data in `predimed.csv` from a large randomimzed, controlled experiment to estimate:

- P(cardiac event | Mediterranean diet)
- P(cardiac event | control diet)

The rule of total probability

Consider the following data on complication rates at a maternity hospital in Cambridge, England:

	Easier deliveries	**Harder deliveries**	**Overall**
Senior doctors	0.052	0.127	0.076
Junior doctors	0.067	0.155	0.072

Would you rather have a junior or senior doctor?

The rule of total probability

Consider the following data on complication rates at a maternity hospital in Cambridge, England:

	Easier deliveries	**Harder deliveries**	**Overall**
Senior doctors	0.052	0.127	0.076
Junior doctors	0.067	0.155	0.072

Would you rather have a junior or senior doctor?

Simpson's paradox. Senior doctors have:

- lower complication rates for easy cases.
- lower complication rates for hard cases.
- higher complication rates overall! (7.6% versus 7.2%.) **Why?**

The rule of total probability

Let's see the table with number of deliveries performed (in parentheses):

	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 rule of total probability

Let's see the table with number of deliveries performed (in parentheses):

	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)

Now we see what's going on:

- 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.

The rule of total probability

It turns out the math of Simpson's paradox can be understood a lot more deeply in terms of something called the *rule of total probability*, or the mixture rule.

This rule sounds fancy, but is actually quite simple.

It says to divide and conquer: **the probability of any event is the sum of the probabilities for all the different ways in which the event can happen.** Really just Kolmogorov's third rule in disguise!

The rule of total probability

Let's see this rule in action for the hospital data.

There are two types of deliveries: easy and hard. So:

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

Now use the rule for conditional probabilities to each joint probability on the right-hand side:

$$\begin{aligned} P(\text{complication}) = {} & P(\text{easy}) \cdot P(\text{complication} \mid \text{easy}) \\ & + P(\text{hard}) \cdot P(\text{complication} \mid \text{hard})\,. \end{aligned}$$

The rule of total probability says that overall probability is a weighted average—a **mixture**—of the two conditional probabilities

The rule of total probability

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 *marginal* or *overall* probabiity of a complication, even though junior doctors have higher *conditional* probabilities of a complication in all scenarios.

Synonyms: overall probability = total probability = marginal probability

The rule of total probability

Here's the formal statement of the rule. Let Ω be any sample space, and let $B_1, B_2, \ldots, B_N$ be a *partition* of Ω—that is, a set of events such that:

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

Now consider any event A. Then

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

Example: drug surveys

Virginia Delaney-Black and her colleagues at Wayne State University gave an anonymous survey to teenagers in Detroit:

- 432 teens were asked whether they had used various drugs.
- Of these 432 teens, 211 agreed to give a hair sample.
- Therefore, for these 211 respondents, the researchers could compare people's answers with an actual drug test.
- Hair samples were analyzed in the aggregate: no hair sample could be traced back to an individual survey or teen.

Citation: 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)

Example: drug surveys

The two sets of results were strikingly different.

- Of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine.
- Bu when the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

Example: drug surveys

The two sets of results were strikingly different.

- Of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine.
- Bu when the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

And the parents lied, too:

- The researchers also asked the parents whether they had used cocaine themselves.
- Only 6.1% said yes.
- But 28.3% of the parents' hair samples came back positive.

Example: drug surveys

Remember:

- these people were guaranteed anonymity
- they wouldn't be arrested or fired for saying admitting drug use
- they willingly agreed to provide a hair sample that could detect drug use.

Yet a big fraction lied about their drug use anyway.

Example: drug surveys

Drug surveys are really important:

- Drug abuse 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! Doctors, schools, and governments all rely on self-reported measures of drug use to guide their thinking on this issue.

Delaney-Black's study asks: **can we trust any of it?**

It's not just drug surveys

Here are some other things that, according to research *on* surveys, people lie about *in* surveys.

Example: a better drug survey

Suppose that you want to learn about the prevalence of drug use among college students. Here's a cute trick that uses probability theory to mitigate someone's incentive to lie.

Suppose that, instead of asking people point-blank, you tell them:

- Flip a coin. Look at the result, but keep it private.
- If the coin comes up heads, please use the space provided to write an answer to question Q1: “Is the last digit of your tax ID number (e.g. SSN) odd?”
- 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?”

Example: a better drug survey

Key fact here: 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.
- The survey designer would never know the difference.

This reduces the incentive to lie.

Let's run the survey! Flip a coin, keep the result private, and then answer the question in public :-)

Analyzing the results

Notation:

- Let Y be the event "a randomly chosen subject answers yes."
- Let Q_1 be the event "the subject answered question 1, about their tax ID number."
- Let Q_2 be the event "the subject answered question 2, about marijuana use."

By the rule of total probability:

$$\begin{aligned} P(Y) &= P(Y, Q_1) + P(Y, Q_2) \\ &= P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2) \end{aligned}$$

Analyzing the results

$$P(Y) = P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2)$$

$P(Y)$ is a weighted average of two conditional probabilities:

- $P(Y \mid Q_1)$, the probability that a subject answers “yes” when answering the tax-ID-number question.
- $P(Y \mid Q_2)$, the probability that a subject answers “yes” when answering the marijuana question.

This equation has five probabilities in it.

- **Which ones do we know from our survey?**
- **Which one do we care about?**

Analyzing the results

Let's solve for $P(Y \mid Q_2)$:

$$P(Y) = P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot \mathbf{P(Y \mid Q_2)}$$

So

$$P(Y \mid Q_2) = \frac{P(Y) - P(Q_1) \cdot P(Y \mid Q_1)}{P(Q_2)}$$

Let's plug in our numbers on the right-hand side and get an answer!

Independence

Two events A and B are *independent* if

$$P(A \mid B) = P(A \mid \text{not } B) = P(A)$$

In words: A and B convey no information about each other:

- P(flip heads second time | flip heads first time) = P(flip heads second time)
- P(stock market up | bird poops on your car) = P(stock market up)
- P(God exists | Longhorns win title) = P(God exists)

So if A and B are independent, then $P(A, B) = P(A) \cdot P(B)$.

Independence

Two events A and B are *conditionally independent*, given C, if

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

A and B convey no information about each other, *once we know C*: $P(A \mid B, C) = P(A \mid C)$.

Neither independence nor conditional independence implies the other.

- It is possible for two outcomes to be *dependent* and yet *conditionally independent.*
- Less intuitively, it is possible for two outcomes to be *independent* and yet *conditionally dependent.*

Independence

Let's see an example. Alice and Brianna live next door to each other and both commute to work on the same metro line.

- A = Alice is late for work.
- B = Brianna is late for work.

A and B are *dependent*: if Brianna is late for work, we might infer that the metro line was delayed or that their neighborhood had bad weather. This means Alice is more likely to be late for work:

$$P(A \mid B) > P(A)$$

Independence

Now let's add some additional information:

- A = Alice is late for work.
- B = Brianna is late for work.
- C = The metro is running on time and the weather is clear.

A and B are *conditionally independent*, given C. If Brianna is late for work but we know that the metro is running on time and the weather is clear, then we don't really learn anything about Alice's commute:

$$P(A \mid B, C) = P(A \mid C)$$

Independence

Same characters, different story:

- A = Alice has blue eyes.
- B = Brianna has blue eyes.

A and B are *independent*: Alice's eye color can't give us information about Brianna's.

Independence

Again, let's add some additional information.

- A = Alice has blue eyes.
- B = Brianna has blue eyes.
- C = Alice and Brianna are sisters.

A and B are *conditionally dependent*, given C: if Alice blue eyes, and we know that Brianna is her sister, then we know something about Brianna's genes. It is now more likely that Brianna has blue eyes.

Checking independence from data

Suppose we have two random outcomes A and B and we want to know if they're independent or not.

Solution:

- Check whether B happening seems to change the probability of A happening.
- That is, verify using data whether $P(A \mid B) = P(A \mid \text{not } B) = P(A)$
- These probabilities won't be *exactly* alike because of statistical fluctuations, especially with small samples.
- But with enough data they should be pretty close if A and B are independent.

Example: the hot hand

NBA Jam c. 1993

Example: the hot hand

The "hot hand hypothesis" says that if a player makes their *previous* shot, they're more likely to make their *next* shot ("He's on fire!"):

$$P(\text{makes next} \mid \text{makes previous}) > P(\text{makes next} \mid \text{misses previous})$$

On the other hand, the "independence hypothesis" says that

$$P(\text{makes next} \mid \text{makes previous}) = P(\text{makes next} \mid \text{misses previous})$$

Example: the hot hand

The next slide show some data on shooting percentages for Dr. J's 1980–81 Philadelphia 76ers.

Key question: do players shoot better, worse, or about the same after they've just *made* a basket, versus how they do after they've just *missed* a basket?

Let's look at the data…

Example: the hot hand

Shooting percentages after:

Player	3 misses	2 misses	1 miss	overall	1 hit	2 hits	3 hits
Julius Erving	0.52	0.51	0.51	0.52	0.52	0.53	0.48
Caldwell Jones	0.50	0.48	0.47	0.43	0.47	0.45	0.27
Maurice Cheeks	0.77	0.6	0.6	0.54	0.56	0.55	0.59
Daryl Dawkins	0.88	0.73	0.71	0.58	0.62	0.57	0.51
Lionel Hollins	0.50	0.49	0.46	0.46	0.46	0.46	0.32
Bobby Jones	0.61	0.58	0.58	0.47	0.54	0.53	0.53
Andrew Toney	0.52	0.53	0.51	0.40	0.46	0.43	0.34
Clint Richardson	0.50	0.47	0.56	0.50	0.50	0.49	0.48
Steve Mix	0.70	0.56	0.52	0.48	0.52	0.51	0.36

Which hypothesis looks right: hot hand or independence?
(Remember small-sample fluctations.)

When independence goes wrong

Suppose we pick a random US family with four male children. What is the probability P that all four will be colorblind?

The probability that a randomly sampled US male is colorblind is about 8%. So the naive answer involves just compounding up this probability:

$$P = 0.08^4 \approx 0.0004$$

What's wrong here?

When independence goes wrong

Colorblindness runs in families (it's an X-linked trait, so males only need one copy on their X chromosome to express the phenotype). So it may be true that

$$P(\text{brother 1 colorblind}) = 0.08$$

But

$$P(\text{brother 2 colorblind} \mid \text{brother 1 colorblind}) = 0.5 \neq 0.08$$

And the same is true for all subsequent brothers: if brother 1 is colorblind, you know that mom is a carrier, and so all her male children have a 50/50 chance of colorblindness (conditional independence, given mom's genes!)

When independence goes wrong

The correct overall probability has to be built up piece by piece using the multiplication rule:

$$
\begin{aligned}
P(\text{brothers 1-4 colorblind}) = {} & P(\text{brother 1 colorblind}) \\
& \times P(\text{brother 2 colorblind} \mid \text{brother 1 colorblind}) \\
& \times P(\text{brother 3 colorblind} \mid \text{brothers 1-2 colorbli} \\
& \times P(\text{brother 4 colorblind} \mid \text{brothers 1-3 colorbli}
\end{aligned}
$$

So:

$$P(\text{brothers 1-4 colorblind}) = 0.08 \times 0.5^3 = 0.01$$

When independence goes wrong

Seems silly, right?

But you'd be surprised at how often people make this mistake! We might call this the “fallacy of mistaken compounding”: assuming events are independent and naively multiplying their probabilities. Out of class, I'm asking you to read two short pieces that illustrate this unfortunate reality:

- How likely is it that birth control could let you down? from the *New York Times*
- An excerpt from Chapter 7 of AIQ: How People and Machines are Smarter Together, by Nick Polson and James Scott.

Checking independence: example

Return to `predimed_intro.R`.

Check the data to answer these questions:

- Are cardiac events independent of sex?
- Are cardiac events conditionally independent of sex, given diet? Hint: estimate P(cardiac event | Mediterranean diet, male) and P(cardiac event | Mediterranean diet, female)
- Are cardiac events conditionally independent of sex, given hypertension status (htn)?

Bayes' Rule

Key fact: all probabilities are contingent on what we know.

When our knowledge changes, our probabilities must change, too.

Bayes' rule tells us how to change them. Suppose A is some event we're interested in and B is some new relevant information. Bayes' rule tells us how to move from a prior probability, $P(A)$, to a posterior probability $P(A \mid B)$ that incorporates our knowledge of B.

Bayes' Rule

$$P(A \mid B) = P(A) \cdot \frac{P(B \mid A)}{P(B)}$$

- $P(A)$ is the prior probability: how probable is A, before having seen data B?
- $P(A \mid B)$ is the posterior probability: how probable is A, now that we've seen data B?
- $P(B \mid A)$ is the likelihood: if A were true, how likely is it that we'd see data B?
- $P(B)$ is the marginal probability of B: how likely is it that we'd see data B overall, regardless of whether A is true or not?

Calculating $P(B)$: use the rule of total probability.

Bayes' Rule: a toy example

Imagine a jar with 1024 normal quarters. Into this jar, a friend places a single two-headed quarter (i.e. with heads on both sides). Your friend shakes the jar to mix up the coins. You draw a single coin at random from the jar, and without examining it closely, flip the coin ten times.

The coin comes up heads all ten times.

Are you holding the two-headed quarter, or an ordinary quarter?

Bayes' Rule: a toy example

Let's see how a posterior probability is calculated using Bayes' rule:

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)} \ .$$

We'll take this equation one piece at a time.

Bayes' Rule: a toy example

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)} .$$

$P(T)$ is the prior probability that you are holding the two-headed quarter.

- There are 1025 quarters in the jar: 1024 ordinary ones, and one two-headed quarter.
- Assuming that your friend mixed the coins in the jar well enough, then you are just as likely to draw one coin as another.
- So $P(T)$ must be 1/1025.

Bayes' Rule: a toy example

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)} \; .$$

Next, what about $P(D \mid T)$, the likelihood of flipping ten heads in a row, given that you chose the two-headed quarter?

- Clearly this is 1.
- If the quarter has two heads, there is no possibility of seeing anything else.

Bayes' Rule: a toy example

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)} \,.$$

Finally, what about $P(D)$, the marginal probability of flipping ten heads in a row? Use the rule of total probability:

$$P(D) = P(T) \cdot P(D \mid T) + P(\text{not } T) \cdot P(D \mid \text{not } T) \,.$$

- $P(T)$ is 1/1025, so $P(\text{not T})$ is 1024/1025.
- $P(D \mid T) = 1$.
- $P(D \mid \text{not T})$ is the probability of a ten-heads "winning streak":

$$P(D \mid \text{not } T) = \left(\frac{1}{2}\right)^{10} = \frac{1}{1024} \,.$$

Bayes' Rule: a toy example

We can now put all these pieces together:

$$
\begin{aligned}
P(T \mid D) &= \frac{P(T) \cdot P(D \mid T)}{P(T) \cdot P(D \mid T) + P(\text{not } T) \cdot P(D \mid \text{not } T)} \\
&= \frac{\frac{1}{1025} \cdot 1}{\frac{1}{1025} \cdot 1 + \frac{1024}{1025} \cdot \frac{1}{1024}} = \frac{1/1025}{2/1025} \\
&= \frac{1}{2}\,.
\end{aligned}
$$

There is only a 50% chance that you are holding the two-headed coin. Yes, flipping ten heads in a row with a normal coin is very unlikely (low likelihood). But so is drawing the one two-headed coin from a jar of 1024 normal coins! (Low prior probability.)

Bayes' Rule in the real world

- Search engines
- Recommender systems
- Medical testing
- Doping control
- Satellite tracking
- Self-driving cars
- Neural decoding in next-gen prostheses. (Longer New Yorker article here.)
- etc!