

# Probability

David Puelz

# Outline

The basics and conditional probability

Independence

Paradoxes, mixtures, and the rule of total probability

# What is probability?

- A measure of **uncertainty**
- Answering the question: “How likely is a given event?”
- As with any mathematical concept, there are a set of **axioms** setting the “ground rules”
- Separately, there are different ways to interpret probability ...
  - (i) **frequentist**: limit of relative frequency after repeating an experiment an infinite number of times (coin flip!)
  - (ii) **Bayesian**: subjective belief about the likelihood of an event occurrence

## Probability basics

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

- $P(\text{coin lands heads}) = 0.5$
- $P(\text{rainy day in Ireland}) = 0.85$
- $P(\text{cold day in Hell}) = 0.0000001$

And so on...

## Probability basics

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

- $P(\text{newborn baby is a boy}) = \frac{106}{206}$
- $P(\text{death due to car accident}) = \frac{11}{100,000}$
- $P(\text{death due to any cause}) = 1$

## Probability basics

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

- $P(\text{newborn baby is a boy}) = \frac{106}{206}$
- $P(\text{death due to car accident}) = \frac{11}{100,000}$
- $P(\text{death due to any cause}) = 1$

Others are synthesized from our best judgments about unique events:

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

## Probability basics: conditioning

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%.”

## Probability basics: conditioning

In statistics, we write this a bit more compactly:

- $P(\text{rain this afternoon} \mid \text{clouds this morning}) = 0.6$
- 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.”

## Probability basics: conditioning

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

- You just gave *Squid Game* a high rating. What's the conditional probability that you will like *Virgin River* or *Love is Blind*?

## Probability basics: conditioning

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

- You just gave *Squid Game* a high rating. What's the conditional probability that you will like *Virgin River* or *Love is Blind*?
- You just bought organic dog food on Amazon. What's the conditional probability that you will also buy a GPS-enabled dog collar?

## Probability basics: conditioning

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

- You just gave *Squid Game* a high rating. What's the conditional probability that you will like *Virgin River* or *Love is Blind*?
- 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 Gavin Newsom (@gavinnewsom) on Instagram. What's the conditional probability that you will respond to a suggestion to follow Greg Abbott (@governorabbott)?

## Probability basics: conditioning

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

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

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”

## Probability basics: conditioning

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



Clearly  $P(A | B) = 1$ : every NBA player can dribble a basketball!

## Probability basics: conditioning

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



But  $P(B | A)$  is nearly zero!

# Uncertain outcomes and probability models

An **uncertain outcome** (more formally called a “random process”) has two key properties:

1. The set of possible outcomes, called the sample space, *is known* beforehand.
2. The particular outcome that occurs is *not known* beforehand.

We denote the **sample space** as  $\Omega$ , and some particular element of the sample space as  $\omega \in \Omega$

# Uncertain outcomes and probability models

Examples:

1. NBA finals, Golden State vs. Toronto:

$$\Omega = \{4-0, 4-1, 4-2, 4-3, 3-4, 2-4, 1-4, 0-4\}$$

# Uncertain outcomes and probability models

Examples:

1. NBA finals, Golden State vs. Toronto:

$$\Omega = \{4-0, 4-1, 4-2, 4-3, 3-4, 2-4, 1-4, 0-4\}$$

2. Temperature in degrees F in Austin on a random day:

$$\Omega = [10, 115]$$

# Uncertain outcomes and probability models

Examples:

1. NBA finals, Golden State vs. Toronto:

$$\Omega = \{4-0, 4-1, 4-2, 4-3, 3-4, 2-4, 1-4, 0-4\}$$

2. Temperature in degrees F in Austin on a random day:

$$\Omega = [10, 115]$$

3. Number of no-shows on an AA flight from Austin to DFW:

$$\Omega = \{0, 1, 2, \dots, N_{\text{seats}}\}$$

# Uncertain outcomes and probability models

Examples:

1. NBA finals, Golden State vs. Toronto:

$$\Omega = \{4-0, 4-1, 4-2, 4-3, 3-4, 2-4, 1-4, 0-4\}$$

2. Temperature in degrees F in Austin on a random day:

$$\Omega = [10, 115]$$

3. Number of no-shows on an AA flight from Austin to DFW:

$$\Omega = \{0, 1, 2, \dots, N_{\text{seats}}\}$$

4. Poker hand

$$\Omega = \text{all possible five-card deals from a 52-card deck}$$

## Uncertain outcomes and probability

An **event** is a *subset of the sample space*, i.e.  $A \subset \Omega$ . For example:

1. NBA finals, Golden State vs. Toronto. Let  $A$  be the event "Toronto wins". Then

$$A = \{3-4, 2-4, 1-4, 0-4\} \subset \Omega$$

2. Austin weather. Let  $A$  be the event "cooler than 90 degrees". Then

$$A = [10, 90) \subset [10, 115]$$

3. Flight no-shows. Let  $A$  be "more than 5 no shows":

$$A = \{6, 7, 8, \dots, N_{\text{seats}}\}$$

## Axioms of probability (Kolmogorov)

These are the [ground rules!](#)

Consider an uncertain outcome with sample space  $\Omega$ . “Probability”  $P(\cdot)$  is a set function that maps  $\Omega$  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  
$$P(A \cup B) = P(A) + P(B).$$
- 3a. **Finite additivity (general)**: For any sets  $A$  and  $B$ ,  
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  
(bonus: prove this with set theory!)

Not that intuitive! Notice no mention of frequencies...

## Summary of terms

- **Uncertain outcome/“random process”**: we know the possibilities ahead of time, just not the specific one that occurs
- **Sample space**: the set of possible outcomes
- **Event**: a subset of the sample space
- **Probability**: a function that maps events to real numbers and that obeys Kolmogorov’s axioms

OK, so how do we actually *calculate* probabilities?

# Counting!

Suppose our sample space  $\Omega$  is a finite set consisting of  $N$  elements  $\omega_1, \dots, \omega_N$ .

Suppose further that  $P(\omega_i) = 1/N$ : each outcome is equally likely, i.e. we have a discrete uniform distribution over possible outcomes.

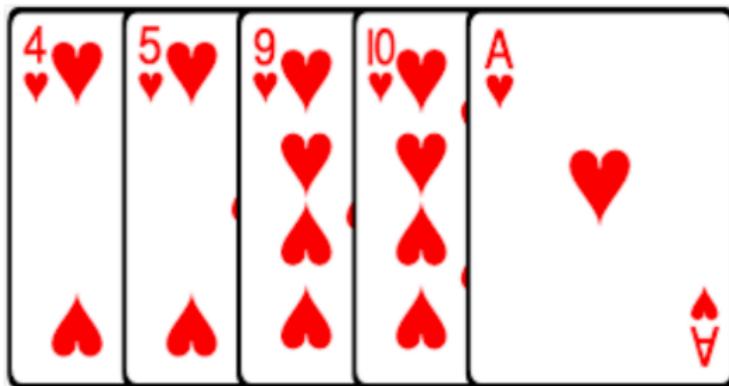
Then for each set  $A \subset \Omega$ ,

$$P(A) = \frac{|A|}{N} = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega}$$

That is, to compute  $P(A)$ , we just need to count how many elements are in  $A$ .

## Counting example

Someone deals you a five-card poker hand from a 52-card deck.  
What is the probability of a flush (all five cards the same suit)?



Note: this is a very historically accurate illustration of probability, given its origins among bored French aristocrats!

## Counting example

- Our sample space has  $N = \binom{52}{5} = 2,598,960$  possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:  
→ There are 13 hearts

## Counting example

- Our sample space has  $N = \binom{52}{5} = 2,598,960$  possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:
  - There are 13 hearts
  - To make a flush with hearts, you need any 5 of these 13 cards.

## Counting example

- Our sample space has  $N = \binom{52}{5} = 2,598,960$  possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:
  - There are 13 hearts
  - To make a flush with hearts, you need any 5 of these 13 cards.
  - Thus there are  $\binom{13}{5} = 1287$  possible flushes with hearts.

## Counting example

- Our sample space has  $N = \binom{52}{5} = 2,598,960$  possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:
  - There are 13 hearts
  - To make a flush with hearts, you need any 5 of these 13 cards.
  - Thus there are  $\binom{13}{5} = 1287$  possible flushes with hearts.
  - The same argument works for all four suits, so there are  $4 \times 1287 = 5,148$  flushes. Thus:

## Counting example

- Our sample space has  $N = \binom{52}{5} = 2,598,960$  possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:
  - There are 13 hearts
  - To make a flush with hearts, you need any 5 of these 13 cards.
  - Thus there are  $\binom{13}{5} = 1287$  possible flushes with hearts.
  - The same argument works for all four suits, so there are  $4 \times 1287 = 5,148$  flushes. Thus:

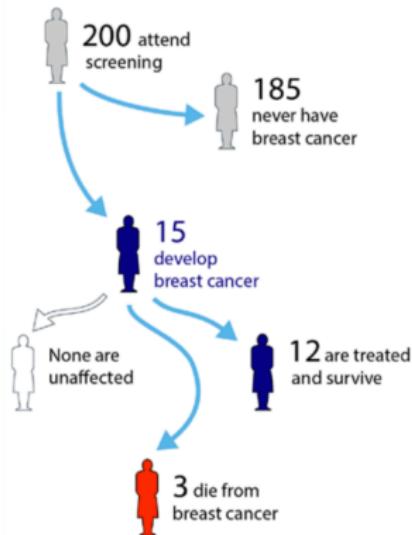
$$P(\text{flush}) = \frac{|A|}{|\Omega|} = \frac{5148}{2598960} = 0.00198079$$

So we know how to count, but what about conditioning?

Probability trees are very useful for this task! This involves counting at different levels of the tree.

# Conditioning example: mammograms

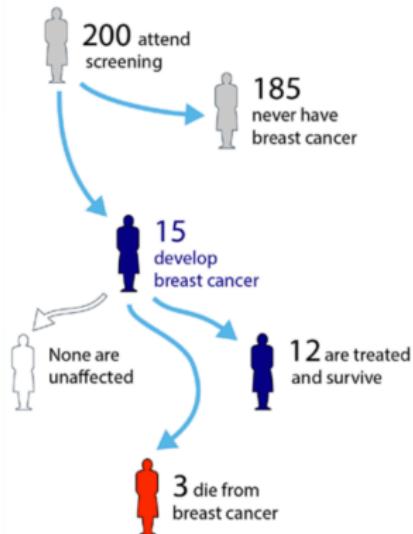
200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) =$
- $P(\text{die, cancer}) =$
- $P(\text{die} \mid \text{cancer}) =$

# Conditioning example: mammograms

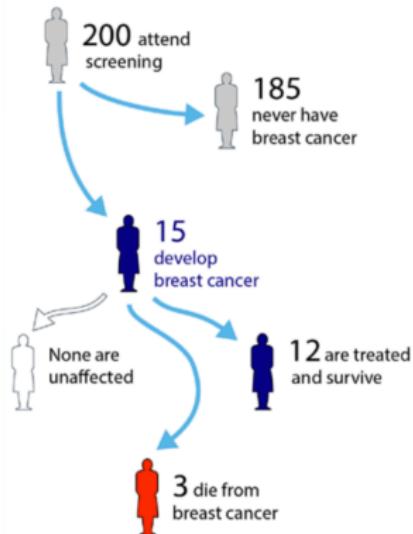
200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- In general, we can estimate the **conditional probability** as:

## Conditioning example: mammograms

200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- In general, we can estimate the **conditional probability** as:

$$P(A | B) = \frac{\text{Frequency of } A \text{ and } B \text{ both happening}}{\text{Frequency of } B \text{ happening}}$$

This is actually a new axiom

**The multiplication rule** – it is an axiom since it can't be derived from the original axioms.

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

## Alternate version

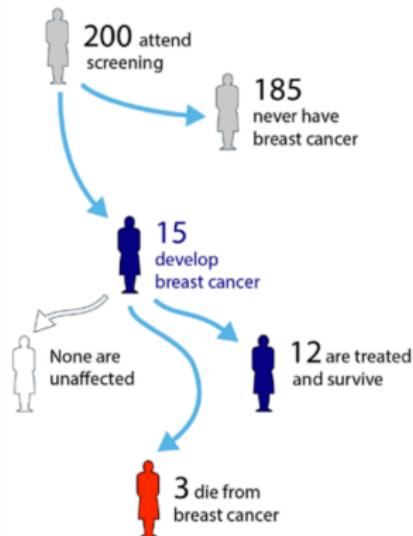
We can also use this alternative version if we want to go in reverse, from a **conditional probability** to a **joint probability**.

It says the same thing with the terms rearranged.

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

# Conditioning example: mammograms (revisited)

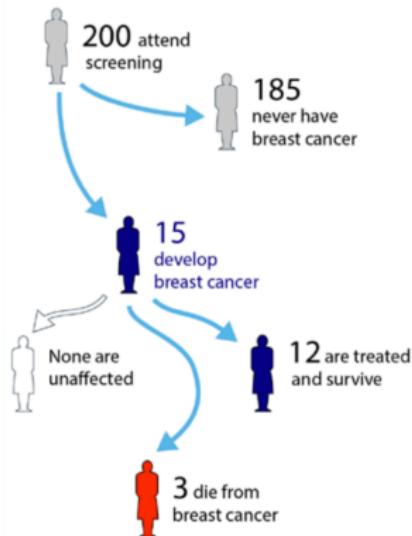
200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- Using the **multiplication rule**, we can estimate the **conditional probability** as:

## Conditioning example: mammograms (revisited)

200 women between 50 and 70  
who attend screening

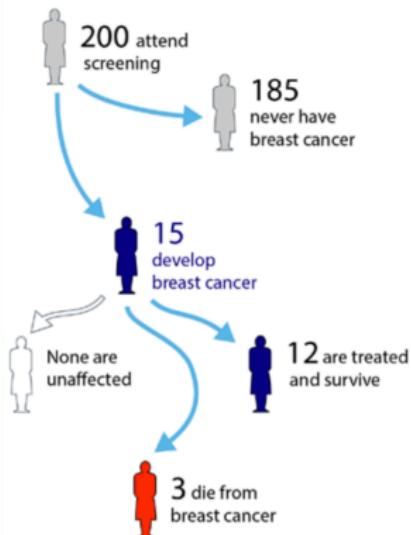


- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- Using the **multiplication rule**, we can estimate the **conditional probability** as:

$$P(\text{die} | \text{cancer}) = \frac{P(\text{die, cancer})}{P(\text{cancer})} = \frac{3/200}{15/200} = \frac{3}{15}$$

# Conditioning example: mammograms (revisited)

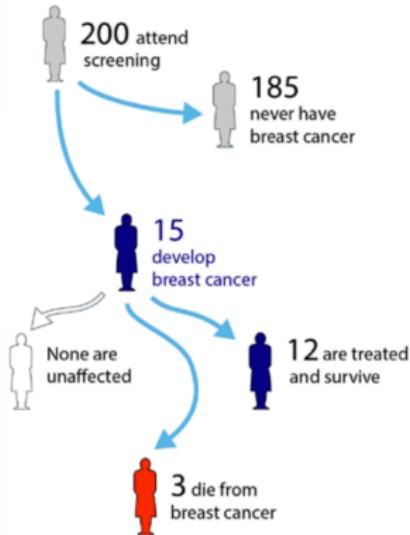
200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- Using the **multiplication rule**, what about computing the **joint probability**?

## Conditioning example: mammograms (revisited)

200 women between 50 and 70  
who attend screening



- $P(\text{cancer}) = \frac{15}{200}$
  - $P(\text{die, cancer}) = \frac{3}{200}$
  - $P(\text{die} | \text{cancer}) = \frac{3}{15}$
- Using the **multiplication rule**, what about computing the **joint probability**?

$$P(\text{die, cancer}) = P(\text{die} | \text{cancer}) \cdot P(\text{cancer}) = \frac{3}{15} \cdot \frac{15}{200} = \frac{3}{200}$$

# Probabilities from contingency tables



# Probabilities from contingency tables



# Probabilities from contingency tables



Suppose you are Netflix ....

You'd like to figure out the chance that **Alistair** will like Saving Private Ryan, given that she likes Band of Brothers.

- What is unknown ( $A$ ): **Alistair** likes Saving Private Ryan
- What is known ( $B$ ): **Alistair** likes Band of Brothers
- **Key question:** What is  $P(A | B)$ ?

Go to the data! (and use the multiplication rule)

<b>Subscriber</b>	<b>Liked SPR?</b>	<b>Liked BoB?</b>
1. Noble Grant	Yes	Yes
2. Jacob Hornstein	No	Yes
3. Gianna Garcia	Yes	No
4. Caterina Erlinger	No	No
5. Max Weinstein	Yes	No
6. Lucy Taylor	Yes	Yes
⋮	⋮	⋮
1575. Ellie Murphy	No	Yes
1576. Kellen Clark	No	No

A nice way to look at this data

(check out the `xtabs()` function in R)

	Liked SPR	Didn't like it
Liked BoB	743	27
Didn't like it	8	798

A nice way to look at this data

(check out the `xtabs()` function in R)

	Liked SPR	Didn't like it
Liked BoB	743	27
Didn't like it	8	798

To figure out *Alistair's* likely preferences:

$$P(\text{Likes SPR} \mid \text{Likes BoB}) = \frac{743}{743 + 27} \approx 0.96$$

A nice way to look at this data

(check out the `xtabs()` function in R)

	Liked SPR	Didn't like it
Liked BoB	743	27
Didn't like it	8	798

To figure out *Alistair's* likely preferences:

$$P(\text{Likes SPR} \mid \text{Likes BoB}) = \frac{743}{743 + 27} \approx 0.96$$

**Q:** What about  $P(\text{Likes BoB} \mid \text{Likes SPR})$ ,  $P(\text{Likes BoB})$ ,  $P(\text{Likes SPR})$ ?

## Conditioning summary

### Moral of the story?

Framing problems in terms of **conditional probabilities** can be immensely useful, whether you are trying to understand individualized preferences or a relationship among uncertain events.

# Independence

Two events  $A$  and  $B$  are **independent** if

$$P(A | B) = P(A)$$

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

- $P(\text{flip heads second time} | \text{flip heads first time}) = P(\text{flip heads second time})$

# Independence

Two events  $A$  and  $B$  are **independent** if

$$P(A | B) = P(A)$$

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

- $P(\text{flip heads second time} | \text{flip heads first time}) = P(\text{flip heads second time})$
- $P(\text{stock market up} | \text{Elon tweets a funny meme}) = P(\text{stock market up})$

# Independence

Two events  $A$  and  $B$  are **independent** if

$$P(A | B) = P(A)$$

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

- $P(\text{flip heads second time} | \text{flip heads first time}) = P(\text{flip heads second time})$
- $P(\text{stock market up} | \text{Elon tweets a funny meme}) = P(\text{stock market up})$
- $P(\text{God exists} | \text{Longhorns win title}) = P(\text{God exists})$

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

# Independence

Independence is often something we *choose to assume* to make probability calculations easier.

# Independence

Independence is often something we *choose to assume* to make probability calculations easier.

In some cases, it is sensible:

- $P(\text{flip 1 heads, flip 2 heads}) = P(\text{flip 1 heads}) \cdot P(\text{flip 2 heads})$
- $P(\text{AAPL up today, AAPL up tomorrow}) = P(\text{AAPL up today}) \cdot P(\text{AAPL up tomorrow})$

# Independence

Independence is often something we *choose to assume* to make probability calculations easier.

In some cases, it is sensible:

- $P(\text{flip 1 heads, flip 2 heads}) = P(\text{flip 1 heads}) \cdot P(\text{flip 2 heads})$
- $P(\text{AAPL up today, AAPL up tomorrow}) = P(\text{AAPL up today}) \cdot P(\text{AAPL up tomorrow})$

In other cases, it is **not** sensible:

- $P(\text{rain, windy}) \neq P(\text{rain}) \cdot P(\text{windy})$
- $P(\text{sibling 1 colorblind, sibling 2 colorblind}) \neq P(\text{sibling 1 colorblind}) \cdot P(\text{sibling 2 colorblind})$

## Conditional independence

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

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

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

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.

## Conditional 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, so in terms of conditional probabilities:

$$P(A | B) > P(A)$$

## Conditional 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:

$$P(A | B, C) = P(A | C)$$

## Conditional 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.

## Conditional independence

Again, let's add some additional information.

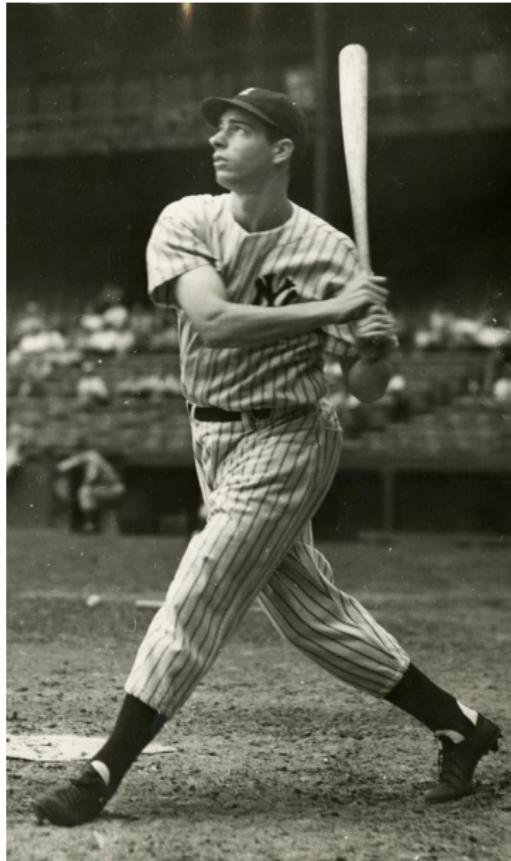
$A$  = Alice has blue eyes.

$B$  = Brianna has blue eyes.

$C$  = Alice and Brianna are sisters.

$A$  and  $B$  are **conditionally dependent**, given  $C$ : if Alice has 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.

Independence  $\iff$  ease of calculation



Independence  $\iff$  ease of calculation

Independence (or conditional independence) is often something we *choose to assume* for the purpose of making calculations easier.

**Example:**

Joe DiMaggio got a hit in about 80% of the baseball games he played in.

Suppose that successive games are independent: if JD gets a hit today, it doesn't change the probability he's going to get a hit tomorrow.

Then  $P(\text{hit in game 1}, \text{hit in game 2}) = 0.8 \cdot 0.8 = 0.64$ .

Independence  $\iff$  ease of calculation

This works for more than two events. For example, Joe DiMaggio had a 56-game hitting streak in the 1941 baseball season. This was pretty unlikely!!

Independence  $\iff$  ease of calculation

This works for more than two events. For example, Joe DiMaggio had a 56-game hitting streak in the 1941 baseball season. This was pretty unlikely!!

$$\begin{aligned} & P(\text{hit game 1, hit game 2, hit game 3, \dots, hit game 56}) \\ &= P(\text{hit game 1}) \cdot P(\text{hit game 2}) \cdot P(\text{hit game 3}) \cdots P(\text{hit game 56}) \\ &= 0.8 \cdot 0.8 \cdot 0.8 \cdots 0.8 \\ &= 0.8^{56} \\ &\approx \frac{1}{250,000} \end{aligned}$$

This is often called the “**compounding rule**.”

Independence  $\iff$  ease of calculation

Let's compare this with the corresponding probability for Pete Rose, a player who got a hit in 76% of his games. He's only slightly less skillful than DiMaggio! But:

$$\begin{aligned} & P(\text{hit game 1, hit game 2, hit game 3, \dots, hit game 56}) \\ &= 0.76^{56} \\ &\approx \frac{1}{5 \text{ million}} \end{aligned}$$

Small difference in one game, but a **big difference** over the long run.

Independence  $\iff$  ease of calculation

What about an average MLB player who gets a hit in 68% of his games?

$$\begin{aligned} & P(\text{hit game 1, hit game 2, hit game 3, \dots, hit game 56}) \\ &= 0.68^{56} \\ &\approx \frac{1}{2.5 \text{ billion}} \end{aligned}$$

Never gonna happen!

# Independence summary

## Summary:

- Joe DiMaggio: 80% one-game hit probability, 1 in 250,000 streak probability
- Pete Rose: 76% one-game hit probability, 1 in 5 million streak probability
- Average player: 68% one-game hit probability, 1 in 2.5 billion streak probability

A small difference in probabilities becomes an enormous difference over the long term.

# Independence summary

## Summary:

- Joe DiMaggio: 80% one-game hit probability, 1 in 250,000 streak probability
- Pete Rose: 76% one-game hit probability, 1 in 5 million streak probability
- Average player: 68% one-game hit probability, 1 in 2.5 billion streak probability

A small difference in probabilities becomes an enormous difference over the long term.

**Moral of the story:** probability compounds **multiplicatively**, like the interest on your credit cards.

## Independence summary

This is a more general assumption that's used in many contexts:

- A mutual-fund manager outperforms the stock market for 15 years straight.
- A World-War II airman completes 25 combat missions without getting shot down, and gets to go home.
- A retired person successfully takes a shower for 1000 days in a row without slipping.
- A child goes 180 school days, or 1 year, without catching a cold from other kids at school. (Good luck!)

**However**, Many smart folks can make mistakes here .. see the reading on our website about birth control.

## Checking independence from data

Suppose we have two random outcomes  $A$  and  $B$  and we want to know if they're independent or not. **How do we go about this?**

## Checking independence from data

Suppose we have two random outcomes  $A$  and  $B$  and we want to know if they're independent or not. **How do we go about this?**

Solution:

- Check whether  $B$  happening seems to change the probability of  $A$  happening
- That is, verify using data whether  $P(A | B) = P(A)$
- 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.

## Paradoxes, mixtures, and the rule of total probability

## The first paradox

Complication rates across 3,690 deliveries at a large maternity hospital in Cambridge, UK.

	low-risk	high-risk	overall
senior doctor	0.052	0.127	
junior doctor	0.067	0.155	

## The first paradox

Complication rates across 3,690 deliveries at a large maternity hospital in Cambridge, UK.

	low-risk	high-risk	overall
senior doctor	0.052	0.127	<b>0.076</b>
junior doctor	0.067	0.155	<b>0.072</b>

## The first paradox

Complication rates across 3,690 deliveries at a large maternity hospital in Cambridge, UK.

	low-risk	high-risk	overall
senior doctor	0.052	0.127	<b>0.076</b>
junior doctor	0.067	0.155	<b>0.072</b>

Q: What doctor do you want delivering your baby?

## The first paradox

- Senior doctors are ...
  - better at **low-risk**
  - better at **high-risk**
  - yet, worse overall?!
- This is an example of **Simpson's paradox**. How is it possible?

## The second paradox

Ten **richest** states and their 2016 electoral college result

<b>Rank</b>	<b>State</b>	<b>Median income</b>	<b>2016 winner</b>
<b>1</b>	Washington, D.C.	\$85,203	Clinton
<b>2</b>	Maryland	\$83,242	Clinton
<b>3</b>	New Jersey	\$81,740	Clinton
<b>4</b>	Hawaii	\$80,212	Clinton
<b>5</b>	Massachusetts	\$79,835	Clinton
<b>6</b>	Connecticut	\$76,348	Clinton
<b>7</b>	California	\$75,277	Clinton
<b>8</b>	New Hampshire	\$74,991	Clinton
<b>9</b>	Alaska	\$74,346	Trump
<b>10</b>	Washington	\$74,073	Clinton

## The second paradox

Ten **poorest** states and their 2016 electoral college result

<b>Rank</b>	<b>State</b>	<b>Median income</b>	<b>2016 winner</b>
<b>42</b>	Tennessee	\$52,375	Trump
<b>43</b>	South Carolina	\$52,306	Trump
<b>44</b>	Oklahoma	\$51,924	Trump
<b>45</b>	Kentucky	\$50,247	Trump
<b>46</b>	Alabama	\$49,861	Trump
<b>47</b>	Louisiana	\$47,905	Trump
<b>48</b>	New Mexico	\$47,169	Clinton
<b>49</b>	Arkansas	\$47,062	Trump
<b>50</b>	Mississippi	\$44,717	Trump
<b>51</b>	West Virginia	\$44,097	Trump

High-income states vote **blue**  
Low-income states vote **red**

“Farmer, factory workers, truck  
drivers, waitresses...”

vs.

The know-it-alls of Manhattan  
and Malibu ... who lord over  
the peasantry with their fancy  
college degrees

“Average Americans, humble,  
long-suffering, working hard,  
who buy their coffee already  
ground”

vs.

“The wealthy, latte-swilling  
liberal elite”

“Real Americans, with a lawnmower in the garage and a flag on the front stoop”

vs.

“Wealthy condo-dwellers with contempt for those who feel chills up their spines at ‘The Star Spangled Banner’”

**And yet ...**

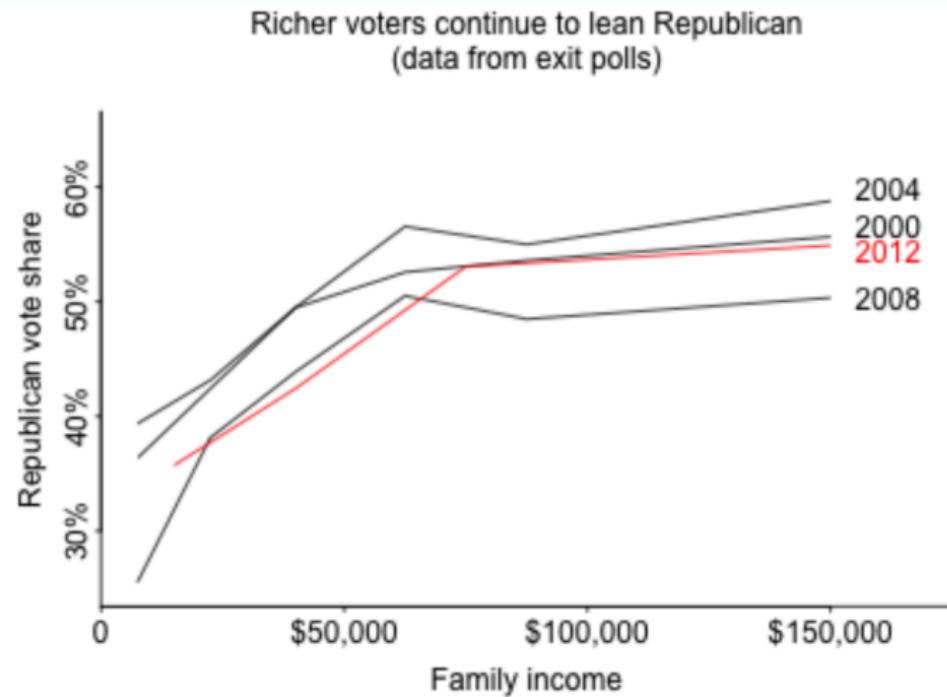
## The second paradox

Presidential vote share by [personal](#) income

	under \$50K		over \$50K	
	Dem.	Rep.	Dem.	Rep.
<b>2004</b>	0.55	0.44	0.43	0.56
<b>2008</b>	0.60	0.38	0.49	0.49
<b>2012</b>	0.54	0.44	0.44	0.54
<b>2016</b>	0.52	0.41	0.47	0.49

## The second paradox

### Presidential vote share by family income



## The second paradox

- For states:
  - higher income means more likely to vote Democrat
  - lower income means more likely to vote Republican
- Yet, for people:
  - higher income means more likely to vote Republican
  - lower income means more likely to vote Democrat
- How is this possible?

## Back to the first paradox

Complication rates and sample sizes across 3,690 deliveries at a large maternity hospital in Cambridge, UK.

	low-risk	high-risk	overall
senior doctor	0.052 (213)	0.127 (102)	0.076 (315)
junior doctor	0.067 (3169)	0.155 (206)	0.072 (3375)

## Rule of total probability

The probability of an event is the sum of the probabilities for all of the different ways that event can happen.

$$P(\text{rain}) = P(\text{rain, wind}) + P(\text{rain, no wind})$$

## Rule of total probability

The probability of an event is the sum of the probabilities for all of the different ways that event can happen.

$$P(\text{rain}) = P(\text{rain, wind}) + P(\text{rain, no wind})$$

$$P(\text{complication}) = P(\text{complication, low-risk}) + P(\text{complication, high-risk})$$

## Rule of total probability

The probability of an event is the sum of the probabilities for all of the different ways that event can happen.

$$P(\text{rain}) = P(\text{rain, wind}) + P(\text{rain, no wind})$$

$$P(\text{complication}) = P(\text{complication, low-risk}) + P(\text{complication, high-risk})$$

Suppose that  $B_1, \dots, B_N$  are mutually exclusive events whose probabilities sum to 1.

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

## Rule of total probability

The probability of an event is the sum of the probabilities for all of the different ways that event can happen.

$$P(\text{rain}) = P(\text{rain, wind}) + P(\text{rain, no wind})$$

$$P(\text{complication}) = P(\text{complication, low-risk}) + P(\text{complication, high-risk})$$

Suppose that  $B_1, \dots, B_N$  are mutually exclusive events whose probabilities sum to 1.

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

Then, for any event  $A$ :

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

## Rule of total probability

	low-risk	high-risk	overall
senior doctor	0.052 (213)	0.127 (102)	<b>0.076 (315)</b>
junior doctor	0.067 (3169)	0.155 (206)	<b>0.072 (3375)</b>

The overall (total) probability of a complication is:

$$P(\text{comp}) = P(\text{comp}, \text{low}) + P(\text{comp}, \text{high})$$

## Rule of total probability

	low-risk	high-risk	overall
senior doctor	0.052 (213)	0.127 (102)	<b>0.076 (315)</b>
junior doctor	0.067 (3169)	0.155 (206)	<b>0.072 (3375)</b>

The overall (total) probability of a complication is:

$$\begin{aligned}P(\text{comp}) &= P(\text{comp}, \text{low}) + P(\text{comp}, \text{high}) \\&= P(\text{low}) \cdot P(\text{comp} | \text{low}) + P(\text{high}) \cdot P(\text{comp} | \text{high})\end{aligned}$$

## Rule of total probability

	low-risk	high-risk	overall
<b>senior doctor</b>	0.052 (213)	0.127 (102)	<b>0.076 (315)</b>
<b>junior doctor</b>	0.067 (3169)	0.155 (206)	<b>0.072 (3375)</b>

The overall (total) probability of a complication:

## Rule of total probability

	low-risk	high-risk	overall
<b>senior doctor</b>	0.052 (213)	0.127 (102)	<b>0.076 (315)</b>
<b>junior doctor</b>	0.067 (3169)	0.155 (206)	<b>0.072 (3375)</b>

The overall (total) probability of a complication:

For senior doctors:

$$P(\text{comp}) = \frac{213}{213+102} \cdot 0.052 + \frac{102}{213+102} \cdot 0.127 = 0.076$$

## Rule of total probability

	low-risk	high-risk	overall
senior doctor	0.052 (213)	0.127 (102)	0.076 (315)
junior doctor	0.067 (3169)	0.155 (206)	0.072 (3375)

The overall (total) probability of a complication:

For senior doctors:

$$P(\text{comp}) = \frac{213}{213+102} \cdot 0.052 + \frac{102}{213+102} \cdot 0.127 = 0.076$$

For junior doctors:

$$P(\text{comp}) = \frac{3169}{3169+206} \cdot 0.067 + \frac{206}{3169+206} \cdot 0.155 = 0.072$$

## First paradox resolved

Senior doctors are...

- better at low-risk *and* high-risk deliveries
- yet worse overall

This is Simpson's paradox in action. Here's what is going on:

- $P(\text{comp} \mid \text{low})$  and  $P(\text{comp} \mid \text{high})$  are both lower for senior doctors
- yet senior doctors **work fewer low-risk cases**:  $P(\text{low})$  is smaller in the mixture!

## First paradox resolved

Senior doctors are...

- better at low-risk *and* high-risk deliveries
- yet worse overall

This is [Simpson's paradox](#) in action. Here's what is going on:

- $P(\text{comp} \mid \text{low})$  and  $P(\text{comp} \mid \text{high})$  are both lower for senior doctors
- yet senior doctors **work fewer low-risk cases**:  $P(\text{low})$  is smaller in the mixture!

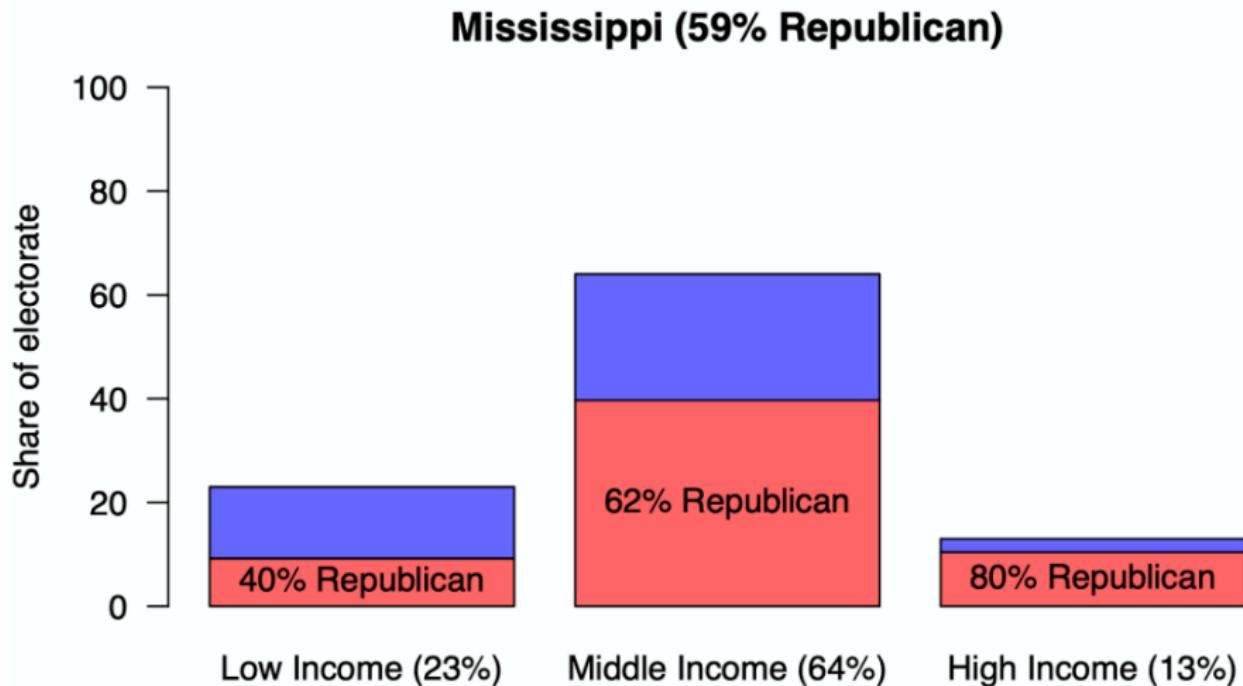
Moral of the story:

- Make sure you're asking the right question
- Always be sensitive to whether probabilities are conditional or unconditional (**marginal**, **total**, **overall**), and which type makes more sense for your situation.

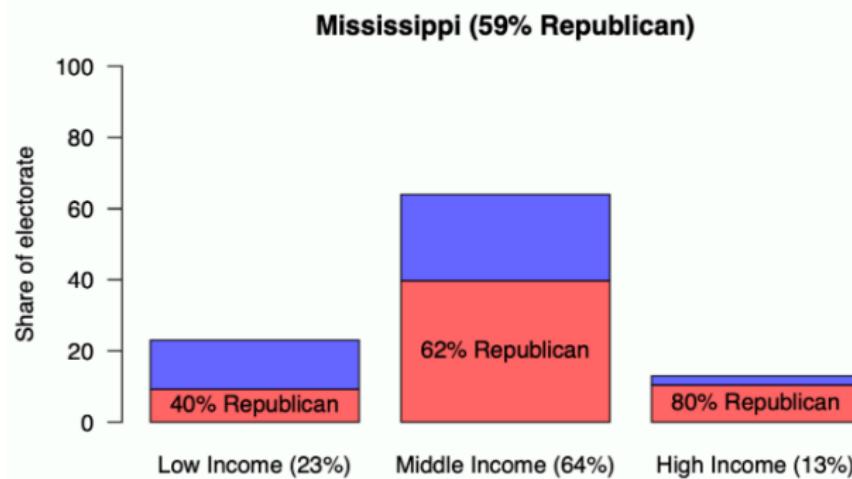
## Back to the second paradox

- For states:
  - higher income means more likely to vote Democrat
  - lower income means more likely to vote Republican
- Yet, for people:
  - higher income means more likely to vote Republican
  - lower income means more likely to vote Democrat
- How is this possible?

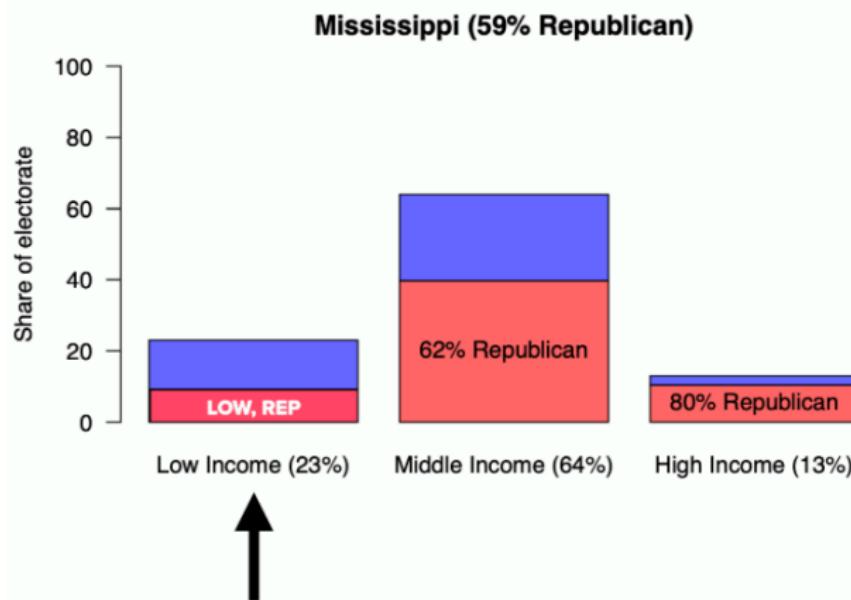
# Law of total probability, Mississippi



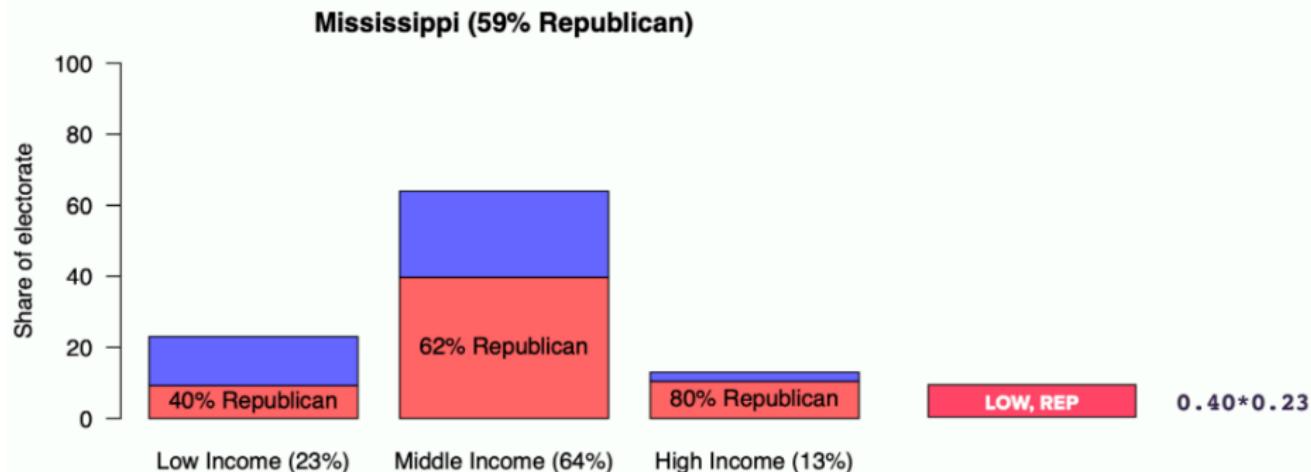
# Law of total probability, Mississippi



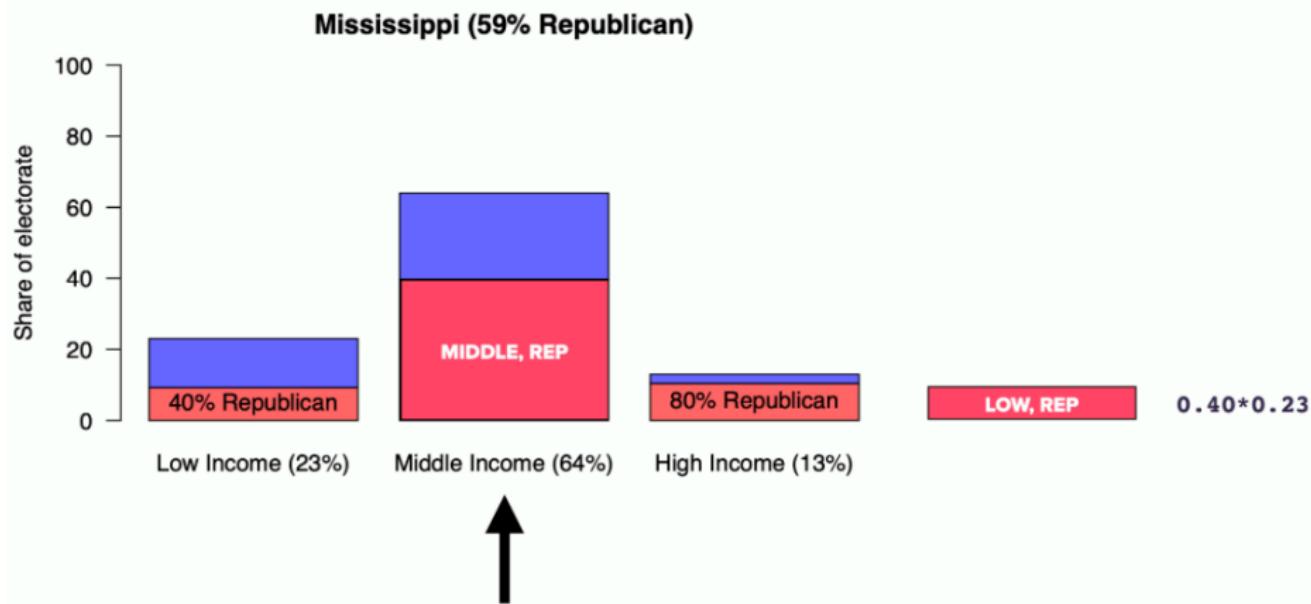
# Law of total probability, Mississippi



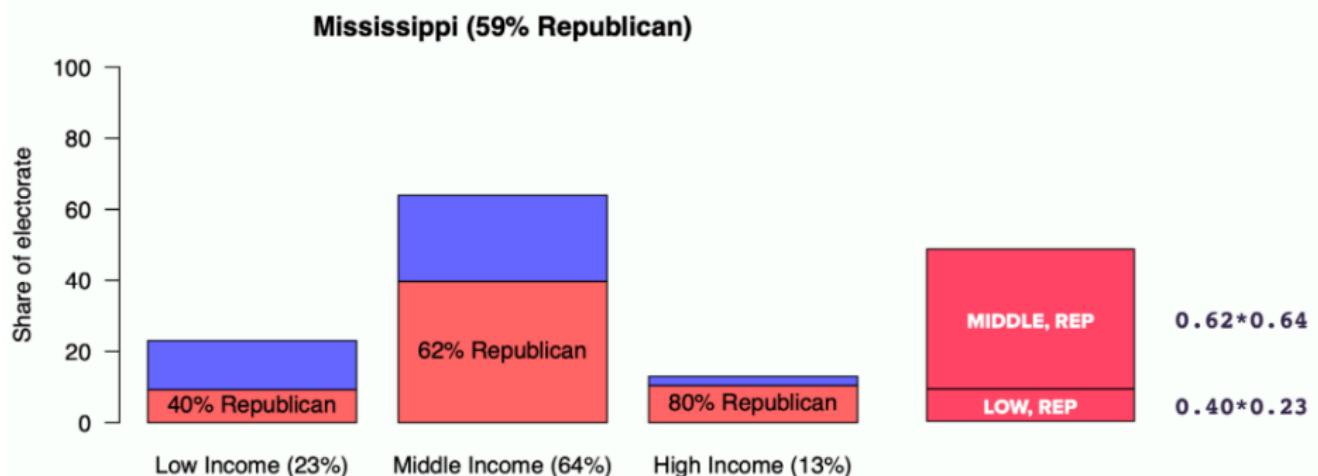
# Law of total probability, Mississippi



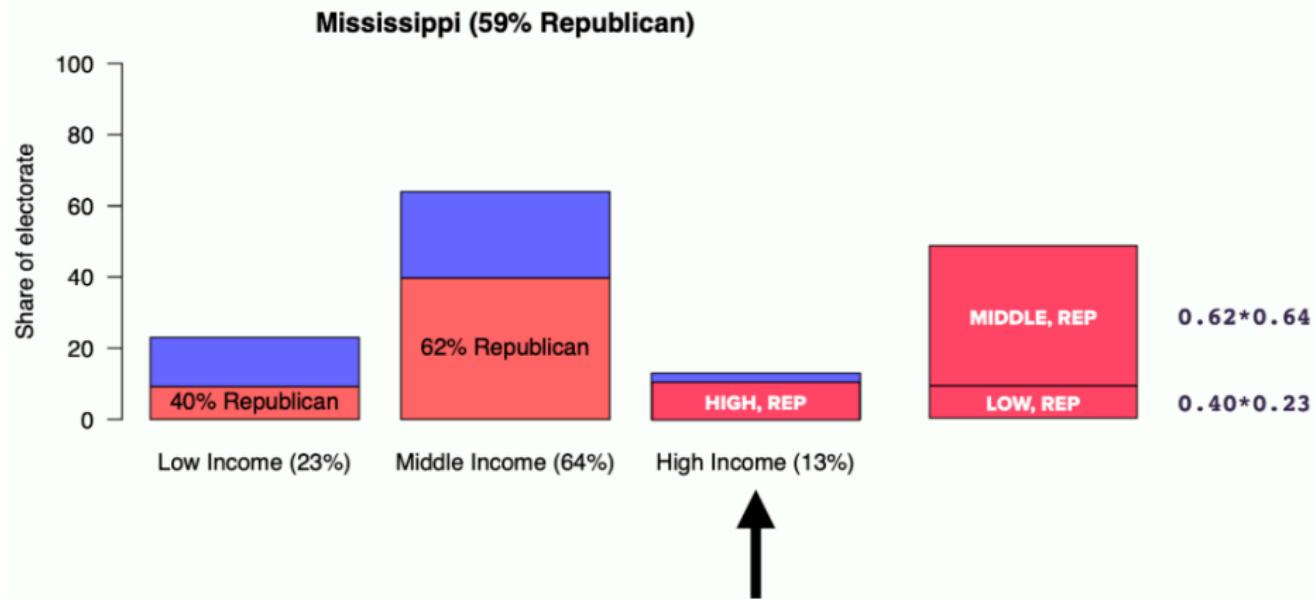
# Law of total probability, Mississippi



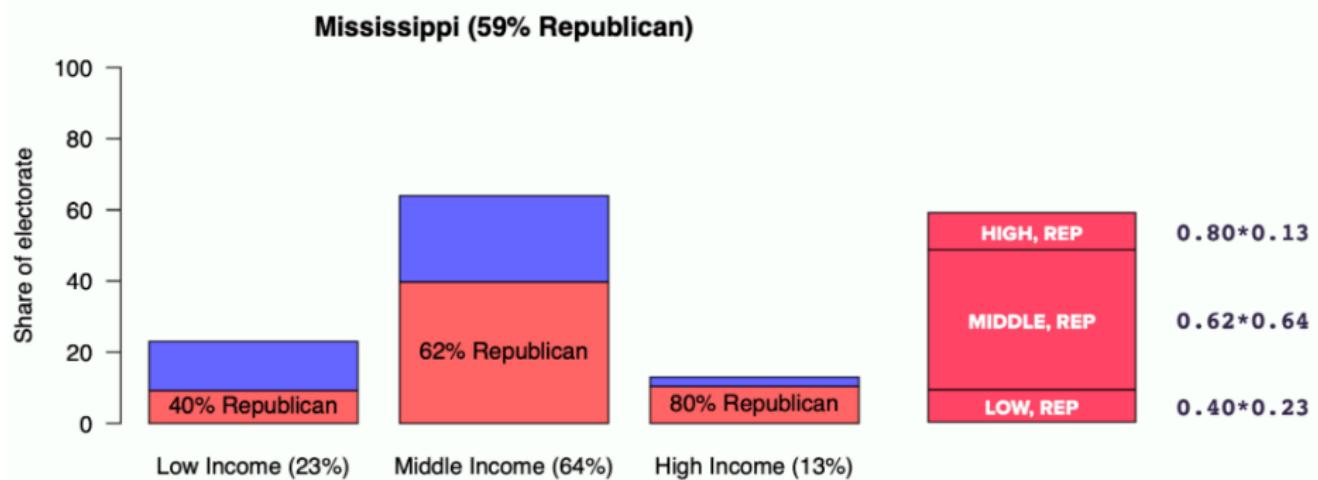
# Law of total probability, Mississippi



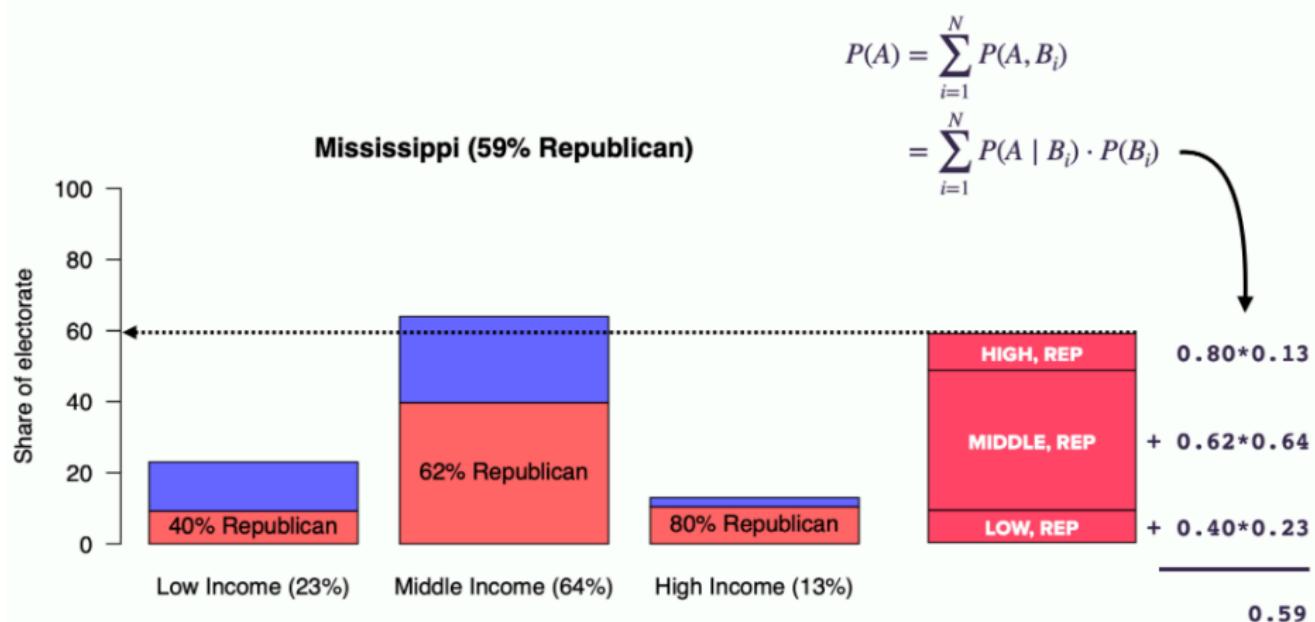
# Law of total probability, Mississippi



# Law of total probability, Mississippi

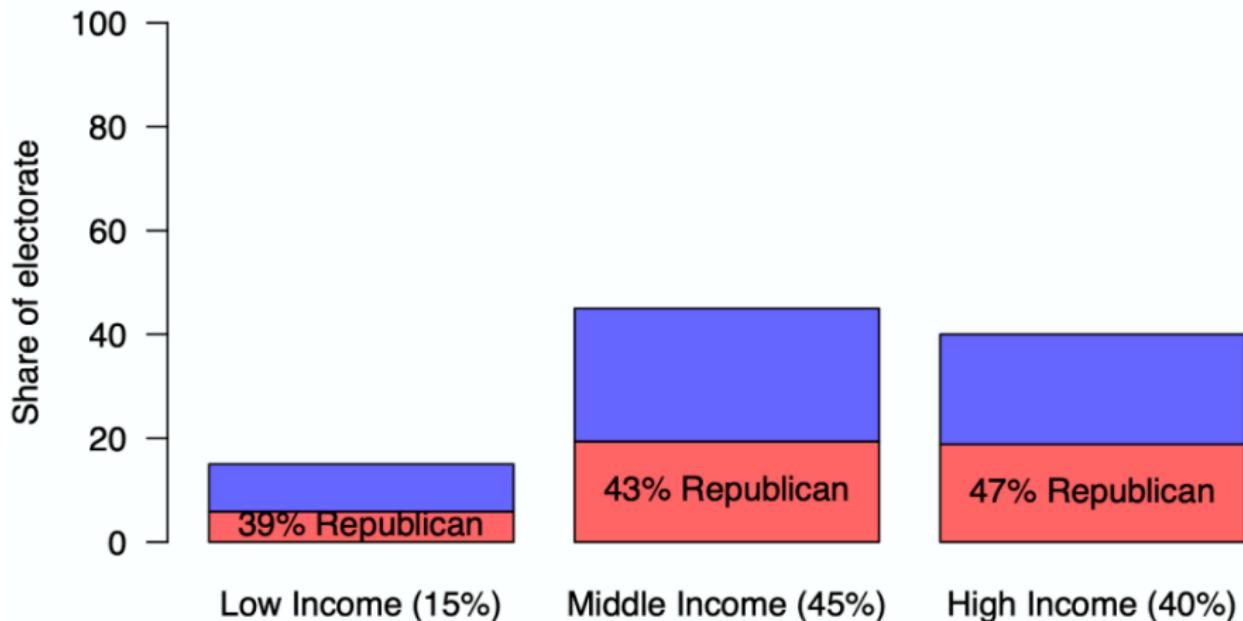


# Law of total probability, Mississippi



And now Connecticut

### Connecticut (44% Republican)



## Connecticut and Mississippi

Here is  $P(\text{Rep} \mid \text{income})$  for each state:

	Low-income	Middle-income	High-income
Connecticut	0.39	0.43	0.47
Mississippi	0.40	0.62	0.80

## Connecticut and Mississippi

Here is  $P(\text{Rep} \mid \text{income})$  for each state:

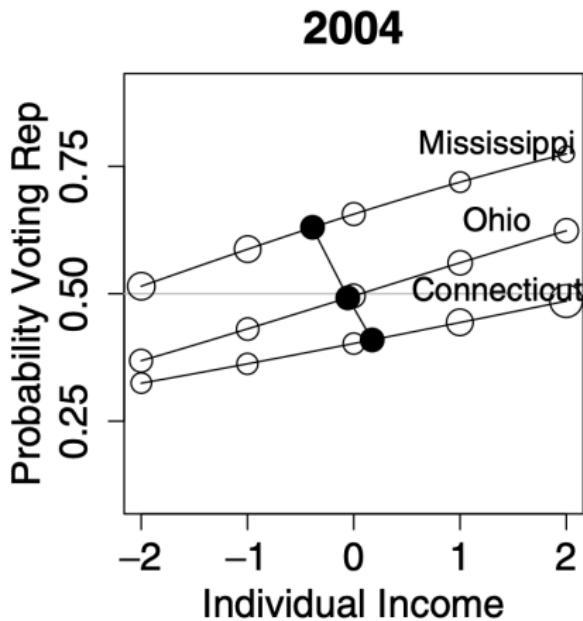
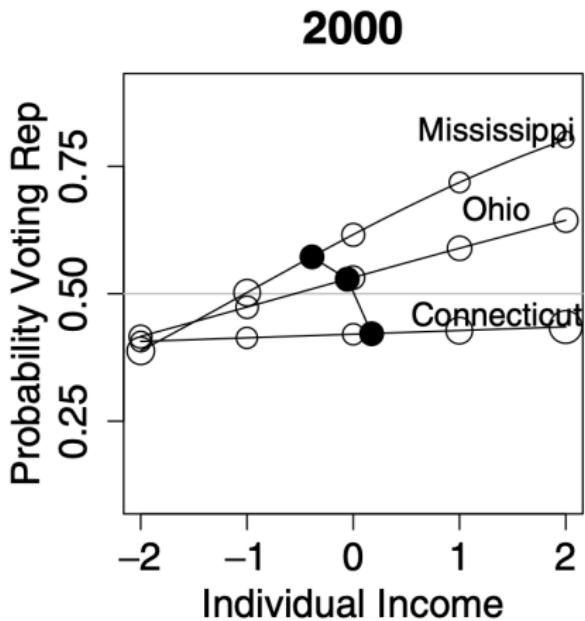
	Low-income	Middle-income	High-income
Connecticut	0.39	0.43	0.47
Mississippi	0.40	0.62	0.80

Q: Does income really tell me anything about why CT is blue and MS is red?

Let's look at Mississippi, Ohio, & Connecticut

(from Gelman et. al., Quarterly Journal of Political Science)

- same story, different election years



Let's look at Mississippi, Ohio, & Connecticut

Paradox 2 resolved, kind of ...

We've seen how, **mechanically**, an individual-level effect can be in one direction, and a group-level effect can be in the other direction.

But, conditioning on income alone **cannot** explain why CT is **blue** and MS is **red**! What can is the relative positioning of the state lines.

What else (other than income) could be driving this relationship?  
(homework)

## The ecological fallacy

**Ecological inference:** looking for associations between cause and effect at the level of groups or populations.

Do groups with higher average levels of A tend to have higher B?

## The ecological fallacy

**Ecological inference:** looking for associations between cause and effect at the level of groups or populations.

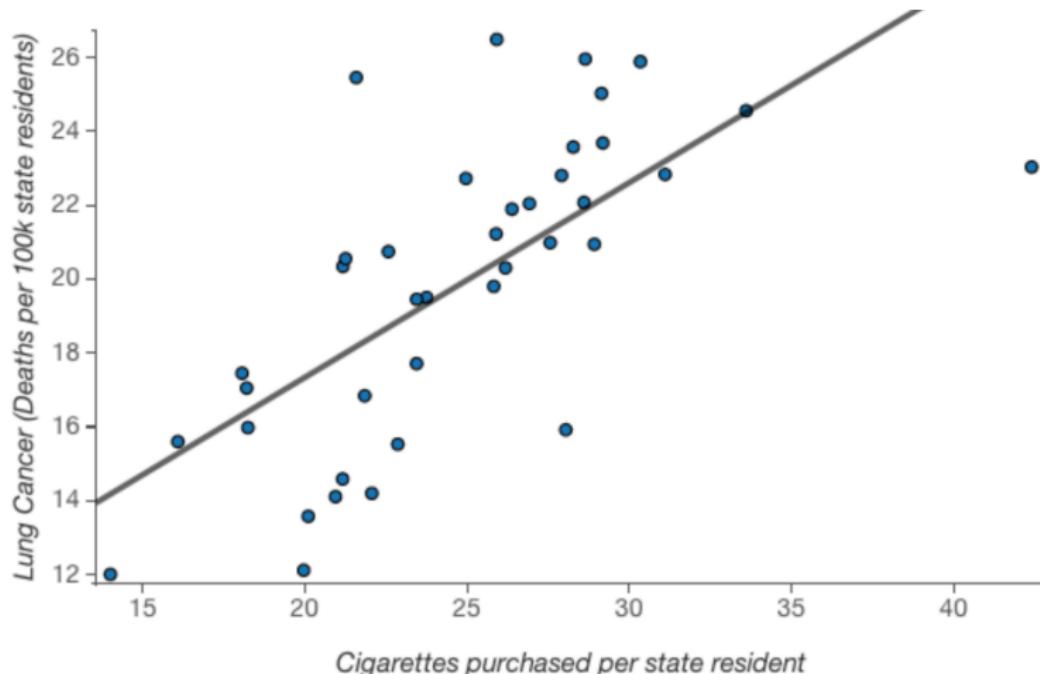
Do groups with higher average levels of A tend to have higher B?

**The ecological fallacy:** assuming, without further justification, that group-level associations accurately reflect individual level associations.

Groups with higher A have higher B, on average. Therefore, individuals with higher A have higher B, on average. ← **not necessarily!!**

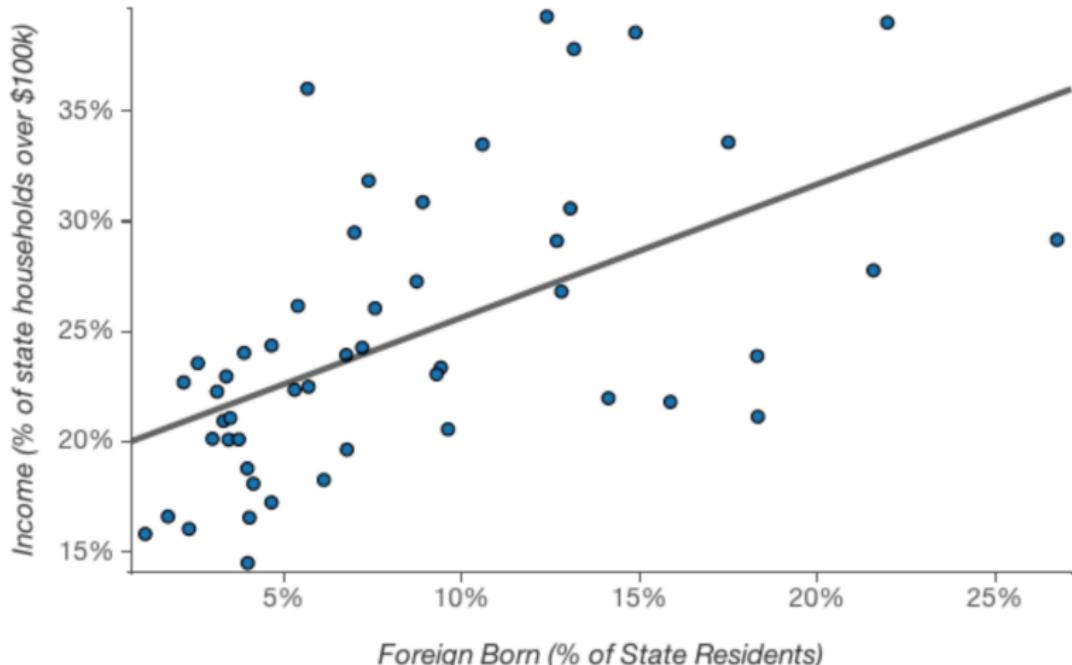
## The ecological fallacy

smoking cigarettes really does increase an individual's risk of lung cancer. This **ecological association** accurately reflects an individual-level trend.



## The ecological fallacy

**... but this one doesn't.** At the individual level, 22.1% of foreign-born residents make more than \$100k, versus 26.1% of US-born residents.



## Take-home messages

- A trend that appears when the data are *separated into individuals/smaller groups* can look different, or even reverse entirely, when the data are *aggregated into larger groups*.

## Take-home messages

- A trend that appears when the data are *separated into individuals/smaller groups* can look different, or even reverse entirely, when the data are *aggregated into larger groups*.
- So what to do? Remember the **rule of total probability!**
  - Pay attention: the level of grouping matters a lot
  - Ask questions: Do we care about a total or conditional probability? Are we missing any lurking variables?
  - Avoid the ecological fallacy: learn to be skeptical when group-level trends are applied to individuals