Probability

David Puelz

Outline

The basics and conditional probability

Independence

Paradoxes, mixtures, and the rule of total probability

Random variables and decision-making

Let's start with a question...

My entire portfolio is in U.S. equities. How would you describe the potential outcomes for my portfolio at the end of the year? What about at the end of 10 years?

Probability and statistics let us talk efficiently about things we are unsure about.

- If I am vaccinated, how likely am I to get COVID-19?
- How much will Amazon sell next quarter?
- What will the return of my retirement portfolio be next year?
- How often will users click on a particular Facebook ad?
- Likelihood of another bank run, chatGPT-like models, etc...

All of these involve inferring or predicting unknown quantities!

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 occurence

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:

- $P(\text{newborn US baby is a boy}) = \frac{98}{198}$
- $P(\text{death due to car accident}) = \frac{11}{100{,}000}$
- $P(\text{death due to any cause}) = 1$

Probability basics

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

- $P(\text{newborn US baby is a boy}) = \frac{98}{198}$
- $P(\text{death due to car accident}) = \frac{11}{100{,}000}$
- $P(\text{death due to any cause}) = 1$

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 *Nailed It* a high rating. What's the conditional probability that you will like *The Terminal List* or *Friends*?

Probability basics: conditioning

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

- You just gave *Nailed It* a high rating. What's the conditional probability that you will like *The Terminal List* or *Friends*?
- 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 *Nailed It* a high rating. What's the conditional probability that you will like *The Terminal List* or *Friends*?
- 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 Quinn Ewers (@quinn_ewers) on Instagram. What's the conditional probability that you will respond to a suggestion to follow Arch Manning (@archmanning)?

Probability basics: conditioning

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"

Probability basics: conditioning

- 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!

Probability basics: conditioning

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

But $P(B \mid A)$ 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 Ω, and some particular element of the sample space as $\omega \in \Omega$

Uncertain outcomes

Examples:

1. NBA finals, Golden State vs. Toronto:

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

Uncertain outcomes

Examples:

1. NBA finals, Golden State vs. Toronto:

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

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

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

Uncertain outcomes

Examples:

1. NBA finals, Golden State vs. Toronto:

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

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:

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

Uncertain outcomes

Examples:

1. NBA finals, Golden State vs. Toronto:

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

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:

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

4. Poker hand

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

Uncertain outcomes

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

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

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":

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

Axioms of probability (Kolmogorov)

These are the ground rules!

Consider an uncertain outcome with sample space Ω. "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 $P(A \cup B) = P(A) + P(B)$.

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 Ω is a finite set consisting of N elements $\omega_1, \ldots, \omega_N$.

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

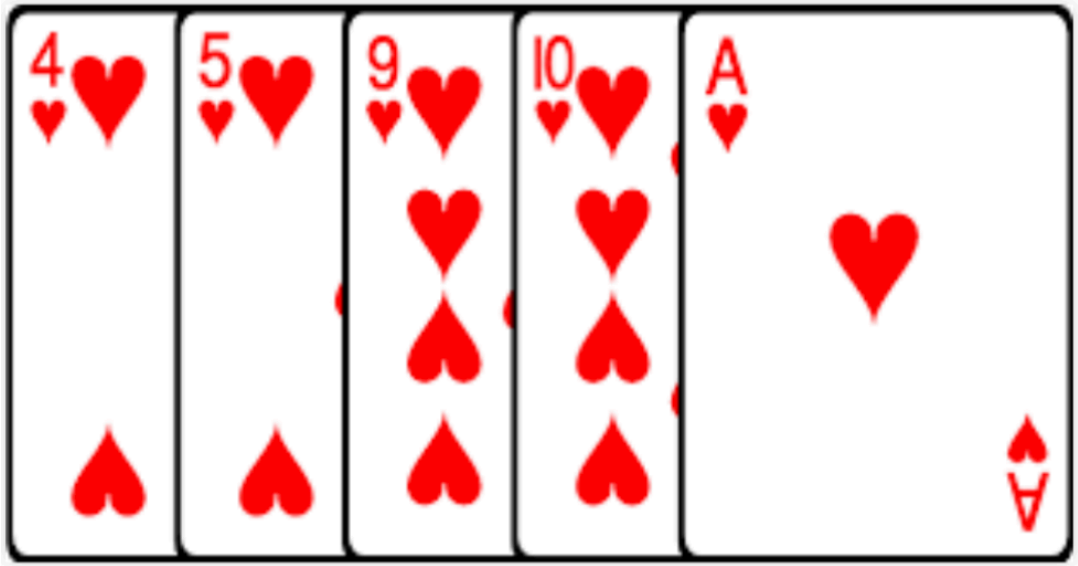
Then for each set $A \subset \Omega$,

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

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:
  $\rightarrow$ 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:
  $\rightarrow$ There are 13 hearts
  $\rightarrow$ 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:
  $\rightarrow$ There are 13 hearts
  $\rightarrow$ To make a flush with hearts, you need any 5 of these 13 cards.
  $\rightarrow$ 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:
  $\rightarrow$ There are 13 hearts
  $\rightarrow$ To make a flush with hearts, you need any 5 of these 13 cards.
  $\rightarrow$ Thus there are $\binom{13}{5} = 1287$ possible flushes with hearts.
  $\rightarrow$ The same argument works for all four suits, so there are 4 $\times$ 1287 = 5,148 flushes. Thus:

Counting example

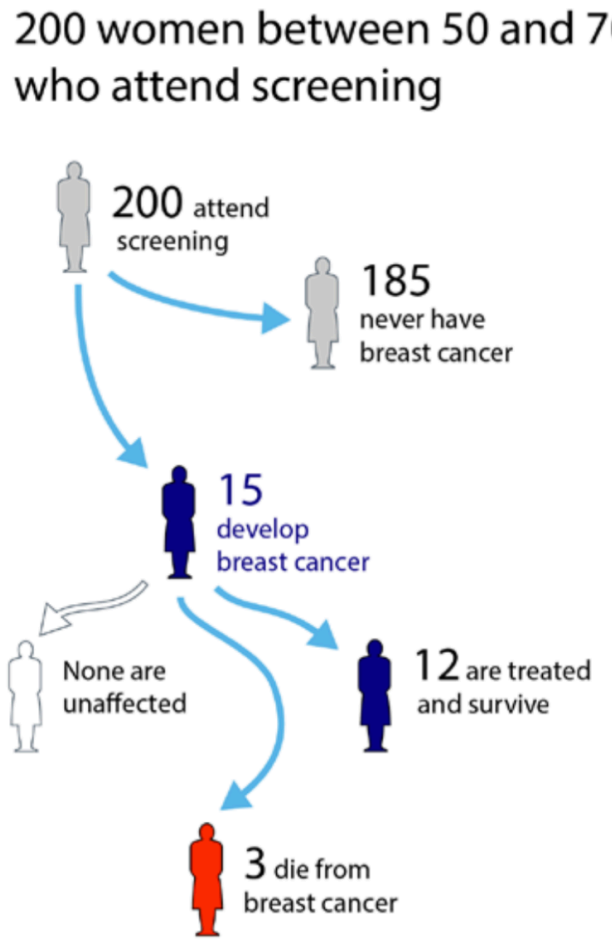
- Our sample space has $N = \binom{52}{5} = 2{,}598{,}960$ possible poker hands, each one equally likely.
- How many possible flushes are there? Let's start with hearts:
  $\rightarrow$ There are 13 hearts
  $\rightarrow$ To make a flush with hearts, you need any 5 of these 13 cards.
  $\rightarrow$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

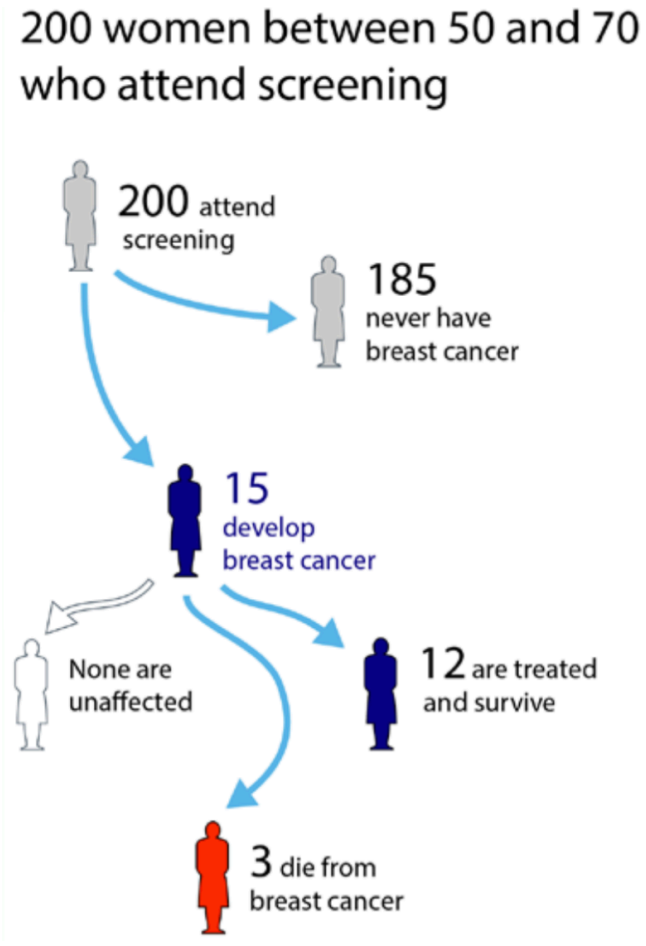
Probability trees are very useful for this task! This involves counting at different levels of the tree.

Conditioning example: Mammograms



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

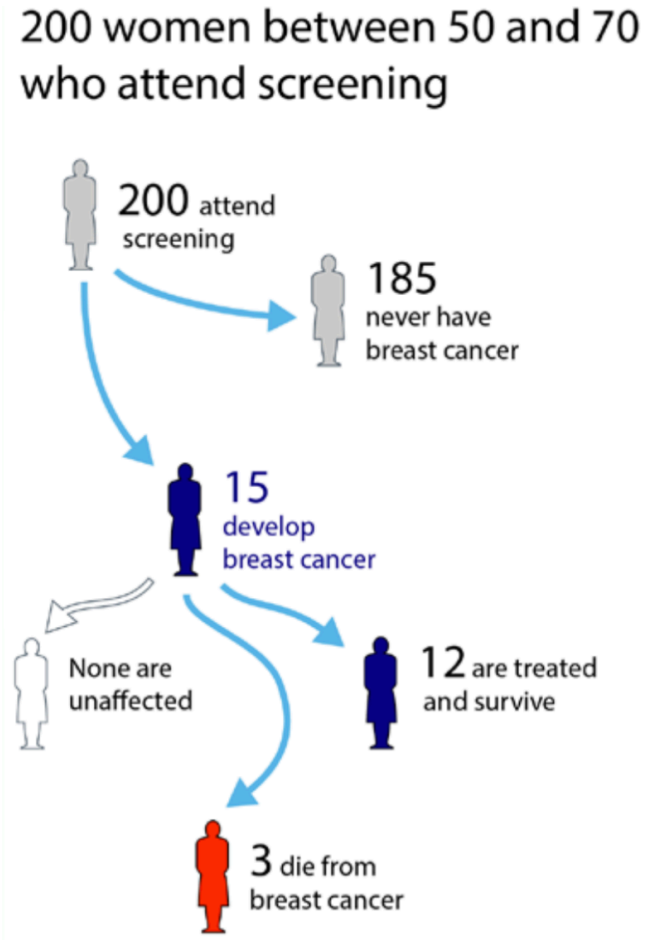
Conditioning example: Mammograms



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

– In general, we can estimate the conditional probability as:

Conditioning example: Mammograms



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

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

This is actually a new axiom

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

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

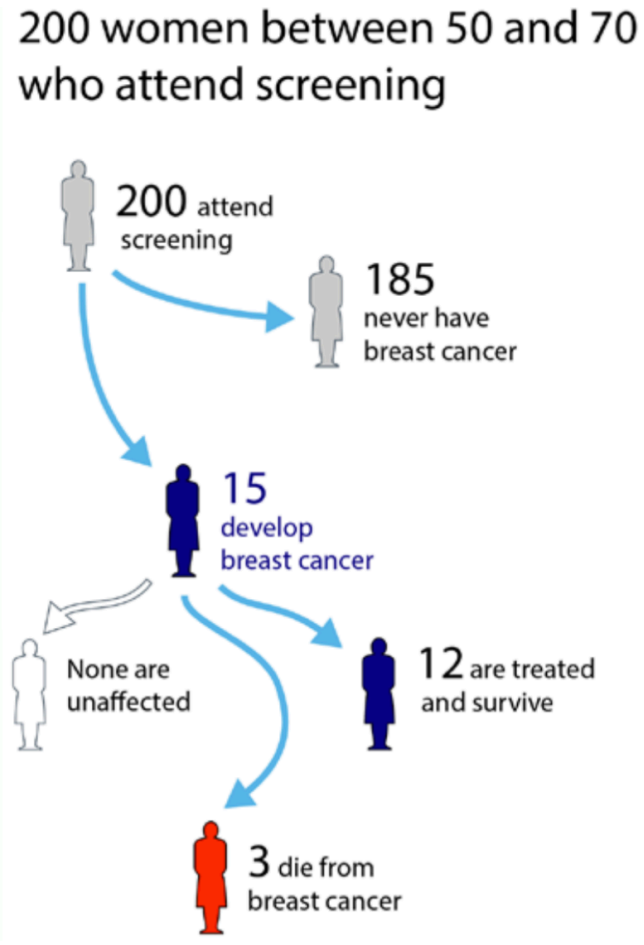
Alternate version

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.

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

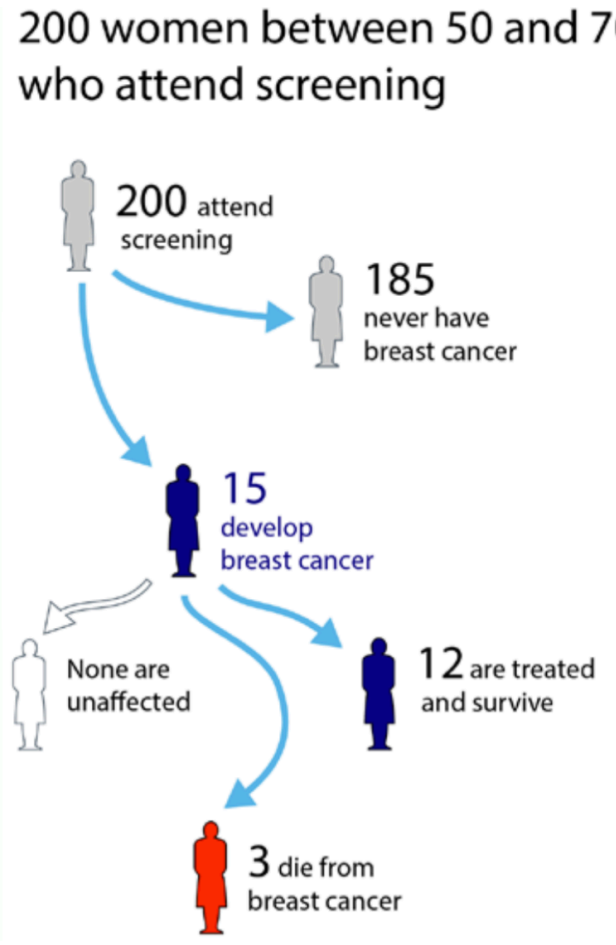
Conditioning example: Mammograms (revisited)



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

– Using the **multiplication rule**, we can estimate the **conditional probability** as:

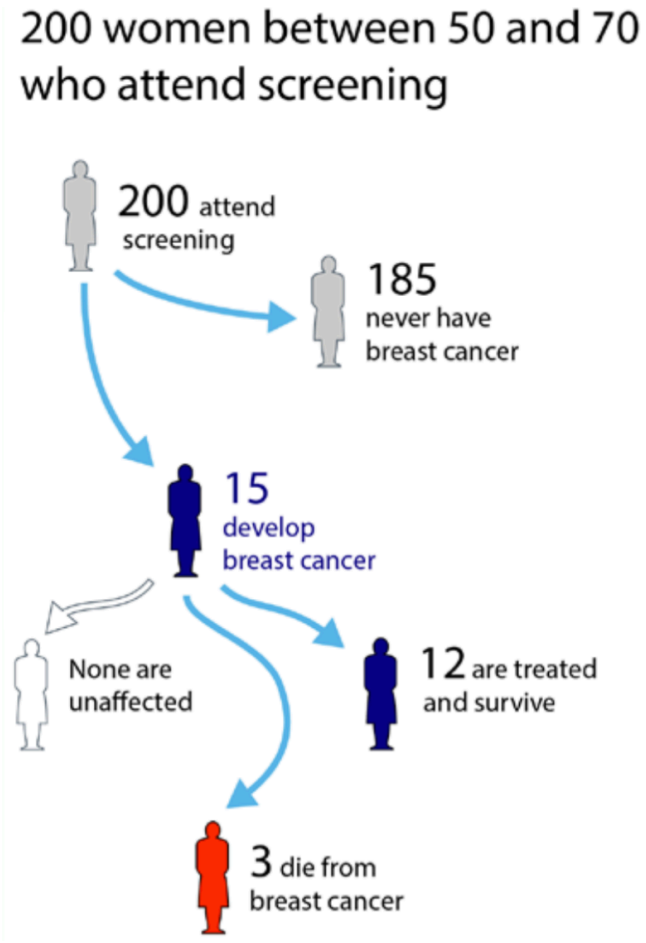
Conditioning example: Mammograms (revisited)



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

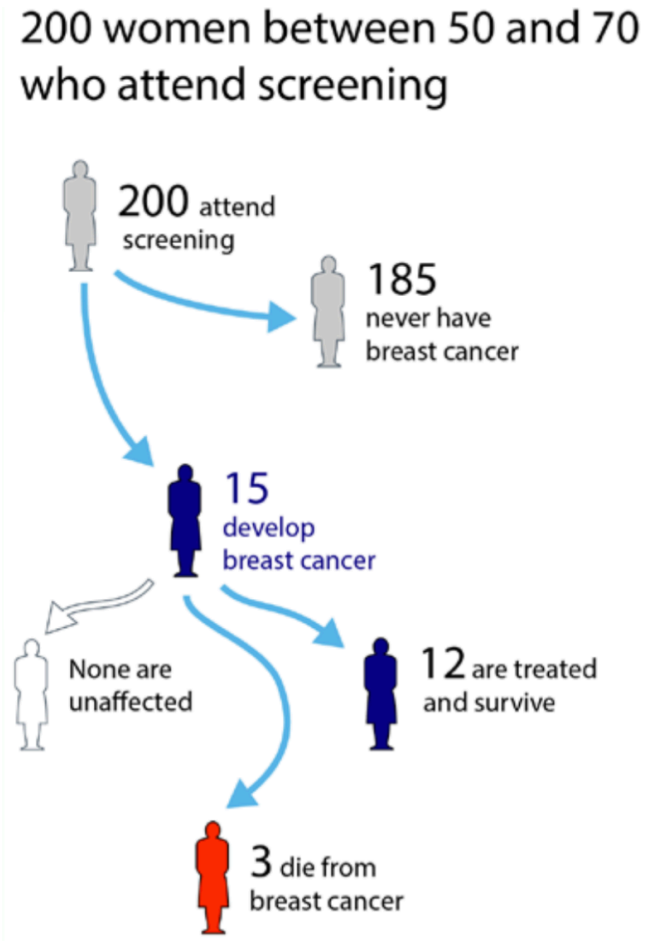
$$P(\text{die} \mid \text{cancer}) = \frac{P(\text{die,cancer})}{P(\text{cancer})} = \frac{3/200}{15/200} = \frac{3}{15}$$

Conditioning example: Mammograms (revisited)



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

Conditioning example: Mammograms (revisited)



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

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

Conditioning example: Sales and the economy

Here's a second example: we often need to answer questions like: **How are my sales impacted by the overall economy?**

Let E denote the performance of the economy next quarter... for simplicity, say $E = 1$ if the economy is expanding and $E = 0$ if the economy is contracting. Let's assume $pr(E = 1) = 0.7$

Conditioning example: Sales and the economy

Let S denote my sales next quarter... and let's suppose the following probability statements:

S	$pr(S\|E=1)$	S	$pr(S\|E=0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

These are our *conditional distributions*

Conditional, Joint and Marginal Distributions

S	$pr(S\|E=1)$	S	$pr(S\|E=0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

- In blue is the conditional distribution of S given $E = 1$
- In red is the conditional distribution of S given $E = 0$
- We read: *the probability of Sales of 4 ($S = 4$)* **given (or conditional on)** *the economy is growing ($E = 1$) is 0.25*

Conditioning example: Sales and the economy

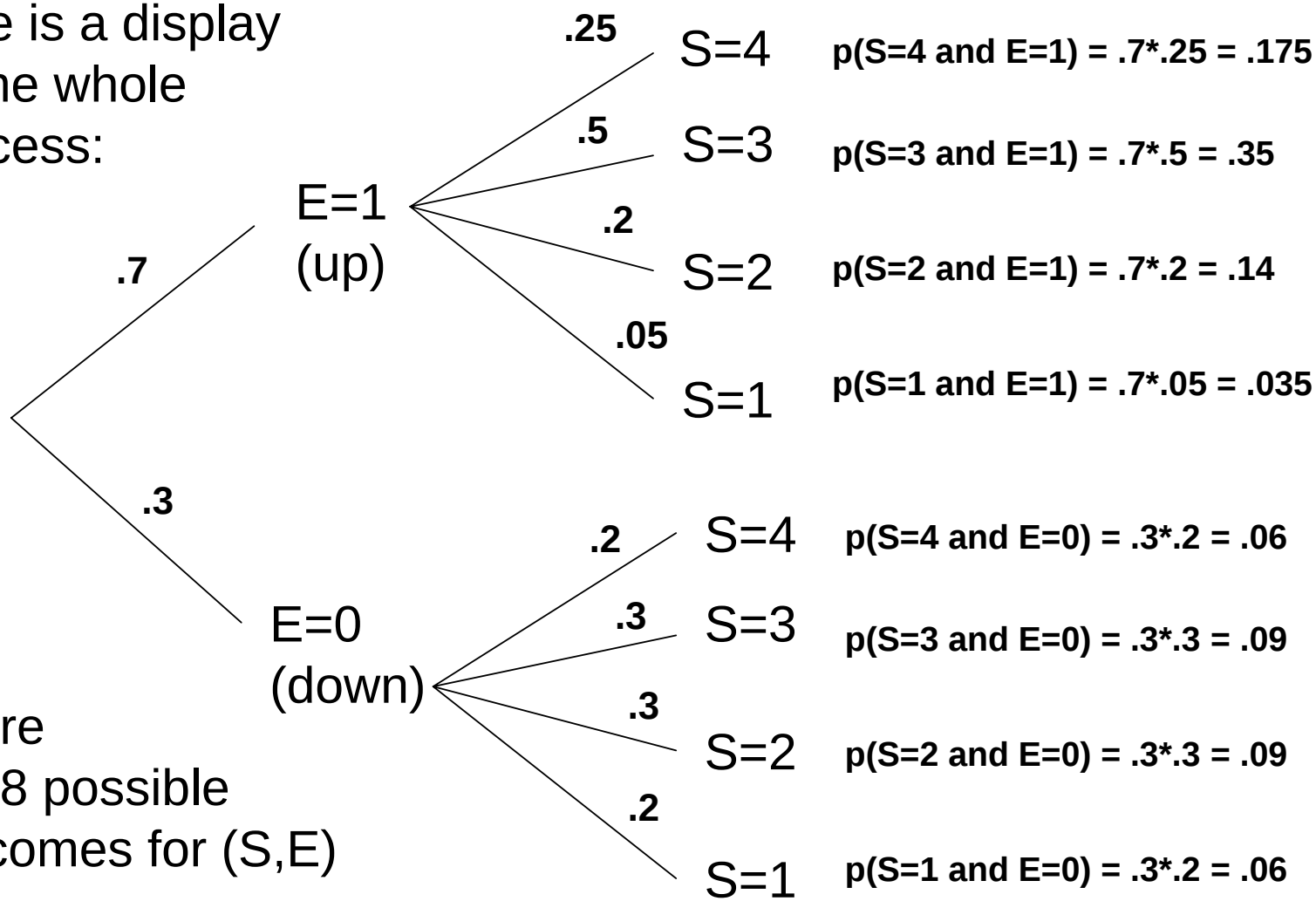
The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

$$\begin{aligned} pr(S = 4 \quad \text{and} \quad E = 1) &= pr(E = 1) \times pr(S = 4|E = 1) \\ &= 0.70 \times 0.25 = 0.175 \end{aligned}$$

In english, 70% of the time the economy grows, and for 1/4 of those times, sales equals 4... "25% of 70% = 17.5%"

Conditioning example: Sales and the economy

here is a display of the whole process:



There are 8 possible outcomes for (S,E)

Conditioning example: Sales and the economy

Why we call marginals marginals... the table represents the joint and at the margins, we get the marginals.

		S = 1	S = 2	S = 3	S = 4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

Conditioning example: Sales and the economy

Example... Given $E = 1$ what is the probability of $S = 4$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$pr(S = 4|E = 1) = \frac{pr(S = 4, E = 1)}{pr(E = 1)} = \frac{0.175}{0.7} = 0.25$$

Conditioning example: Sales and the economy

Example... Given $S = 4$ what is the probability of $E = 1$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$pr(E = 1|S = 4) = \frac{pr(S = 4, E = 1)}{pr(S = 4)} = \frac{0.175}{0.235} = 0.745$$

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 Rose will like Saving Private Ryan, given that she likes Band of Brothers.

- What is unknown (A): Rose likes Saving Private Ryan
- What is known (B): Rose likes Band of Brothers
- Key question: What is $P(A \mid B)$?

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

Subscriber	Liked SPR?	Liked BoB?
1. Zachary Adrian	Yes	Yes
2. Luke Davis	No	Yes
3. Israel Escamilla	Yes	No
4. Jessica Ward	No	No
5. Sarah Ramirtha	Yes	No
6. Emily Carter	Yes	Yes
⋮	⋮	⋮
1575. Seth Friberg	No	Yes
1576. Jiali O'Riain	No	No

A nice way to look at this data

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

A nice way to look at this data

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

To figure out Rose'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

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

To figure out Rose'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

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

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

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

Independence

Two events A and B are independent if

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

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

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

Independence

Two events A and B are independent if

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

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

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

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:

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

Independence

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

In some cases, it is sensible:

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

In other cases, it is not sensible:

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

Conditional independence

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

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

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.

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:

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

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:

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

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.

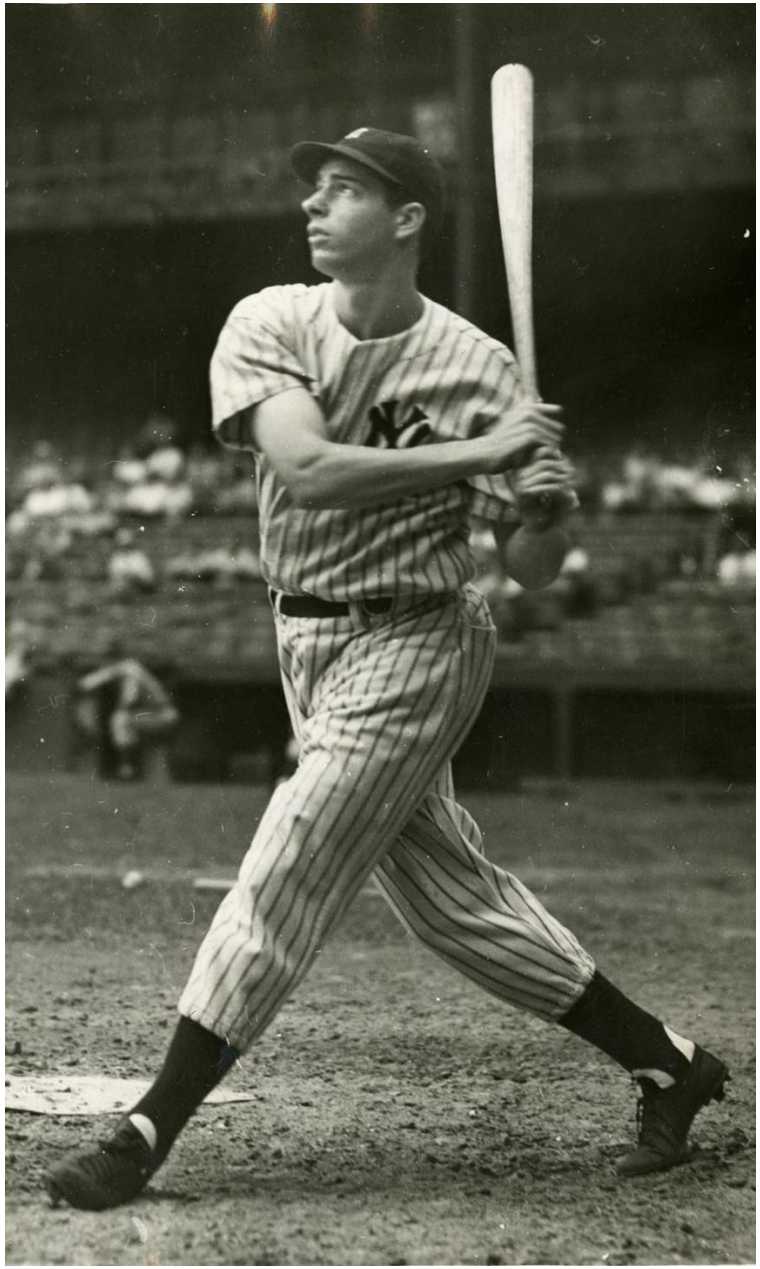
A = Alice has blue eyes.

B = Brianna has blue eyes.

C = Alice and Brianna are sisters.

A and B are conditionally dependent, given C: if Alice 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 $\Longleftrightarrow$ 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!!

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

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:

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

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?

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

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 ..

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 $P(A \mid B) = P(A)$
- 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	**0.076**
junior doctor	0.067	0.155	**0.072**

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	**0.076**
junior doctor	0.067	0.155	**0.072**

Q: What doctor do you want delivering your baby?

The first paradox

- Senior doctors are ...
  $\rightarrow$ better at low-risk
  $\rightarrow$ 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

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

The second paradox

Ten poorest states and their 2016 electoral college result

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

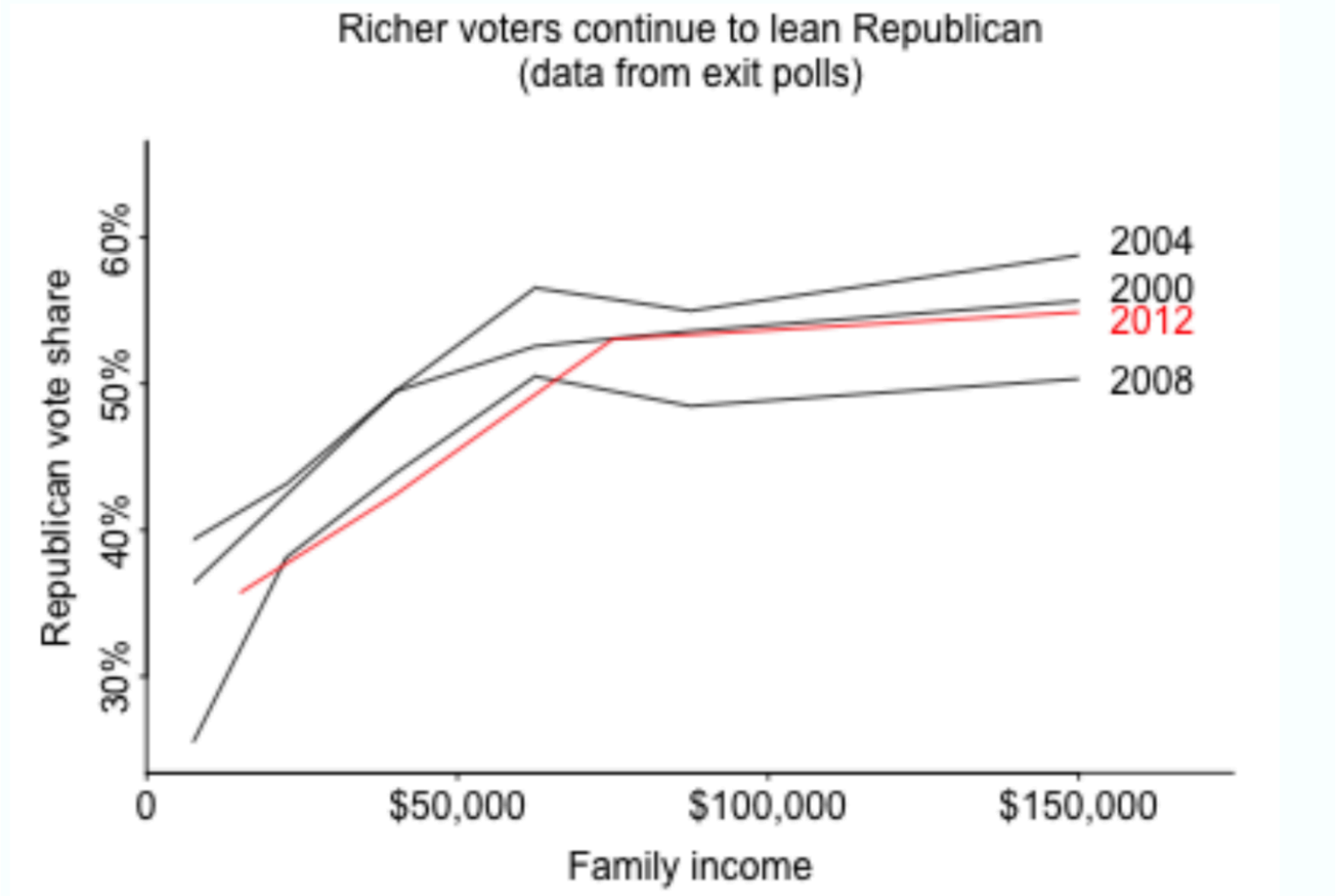
The second paradox

Presidential vote share by personal income

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

The second paradox

Presidential vote share by family income



The second paradox

- For states:
  $\rightarrow$ higher income means more likely to vote Democrat
  $\rightarrow$ lower income means more likely to vote Republican
- Yet, for people:
  $\rightarrow$ higher income means more likely to vote Republican
  $\rightarrow$ 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.

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

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.

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

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

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.

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

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

Suppose that $B_1, \ldots B_N$ 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.

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

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

Suppose that $B_1, \ldots B_N$ 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:

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

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 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)	**0.076** (315)
junior doctor	0.067 (3169)	0.155 (206)	**0.072** (3375)

The overall (total) probability of a complication is:

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

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:

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:

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

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!

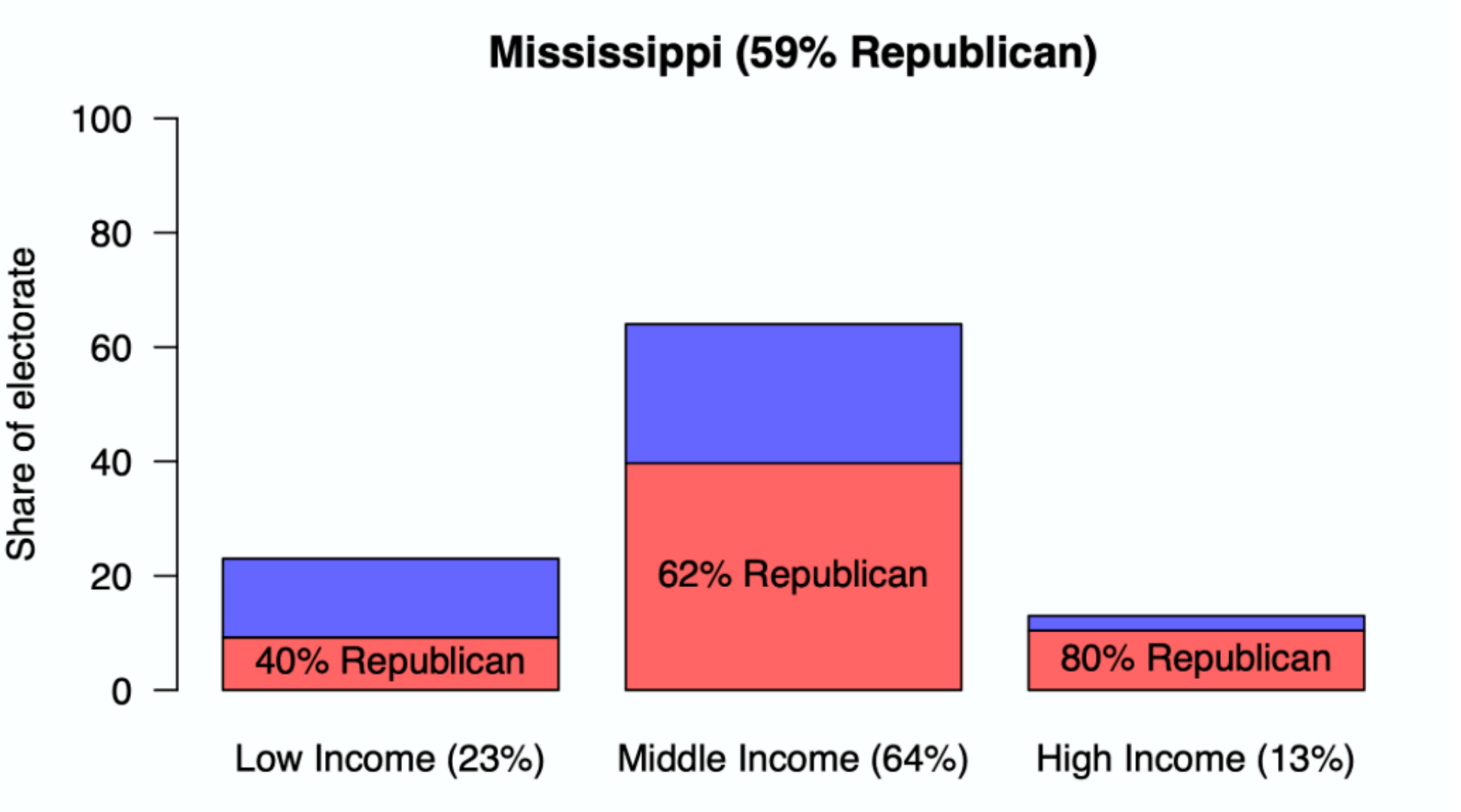
Moral of the story:

- 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.

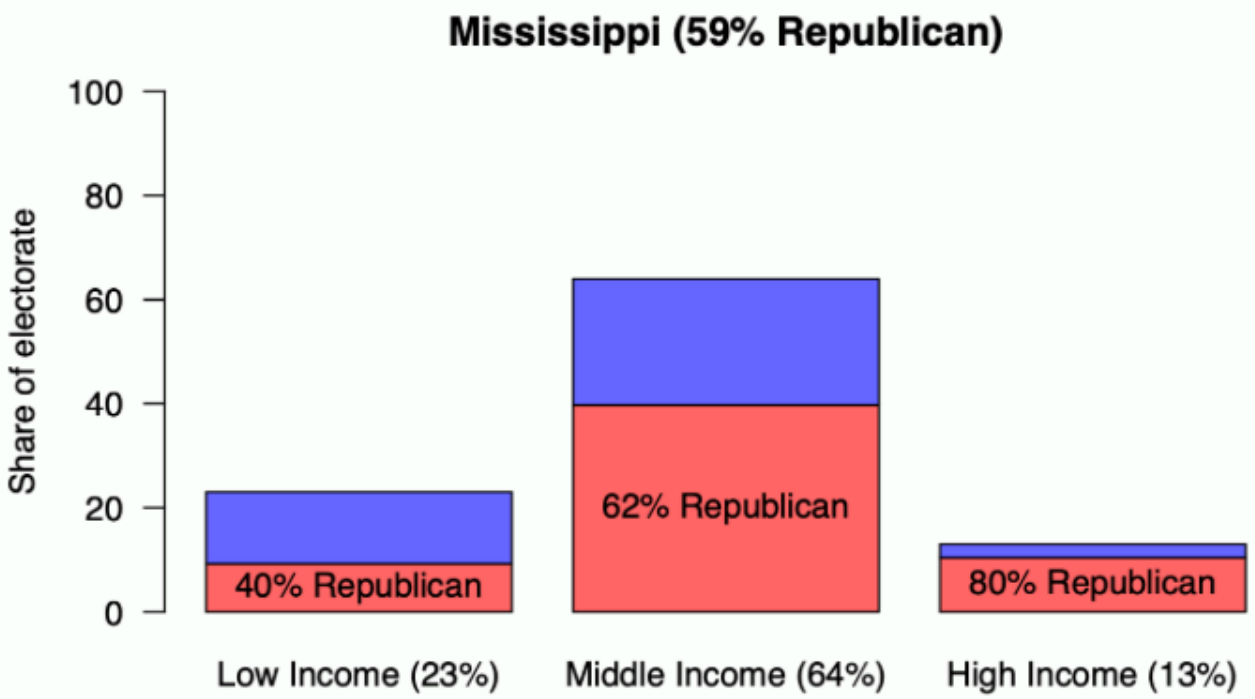
Back to the second paradox

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

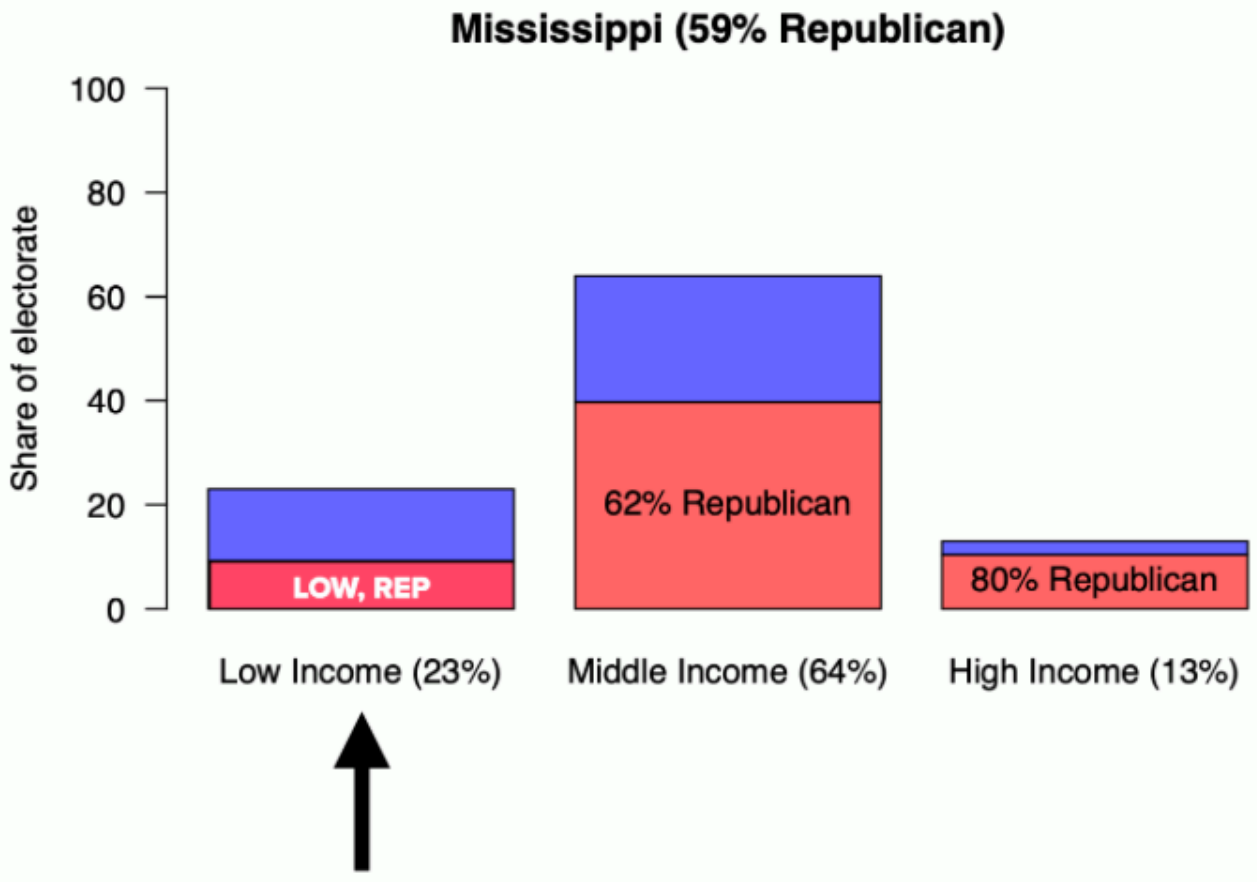
Law of total probability, Mississippi



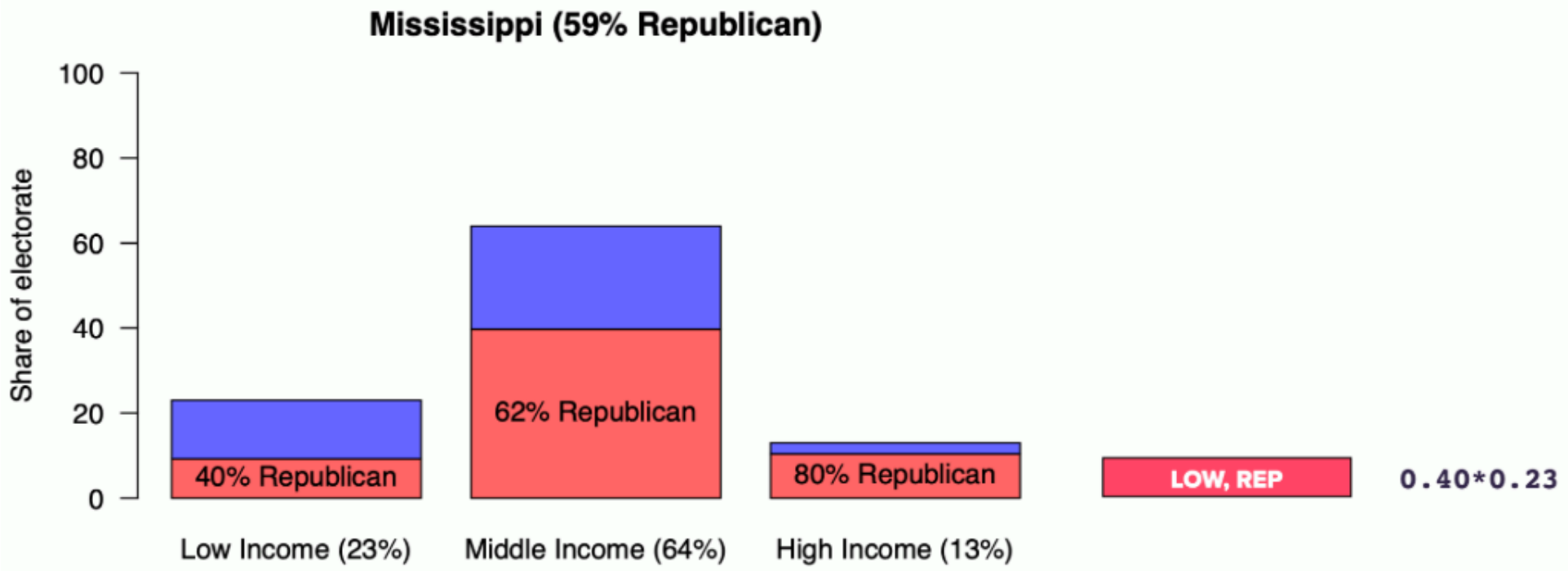
Law of total probability, Mississippi



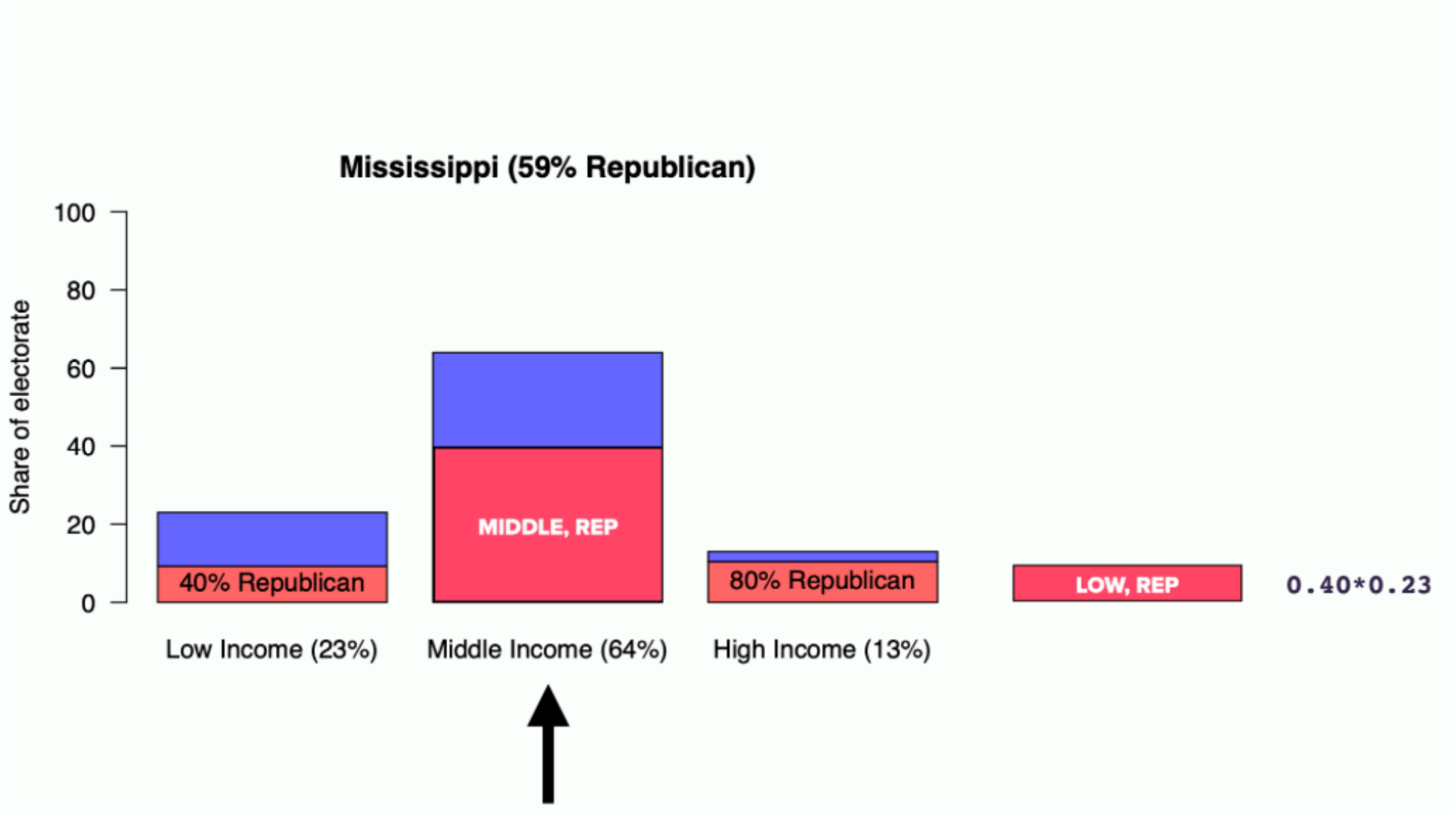
Law of total probability, Mississippi



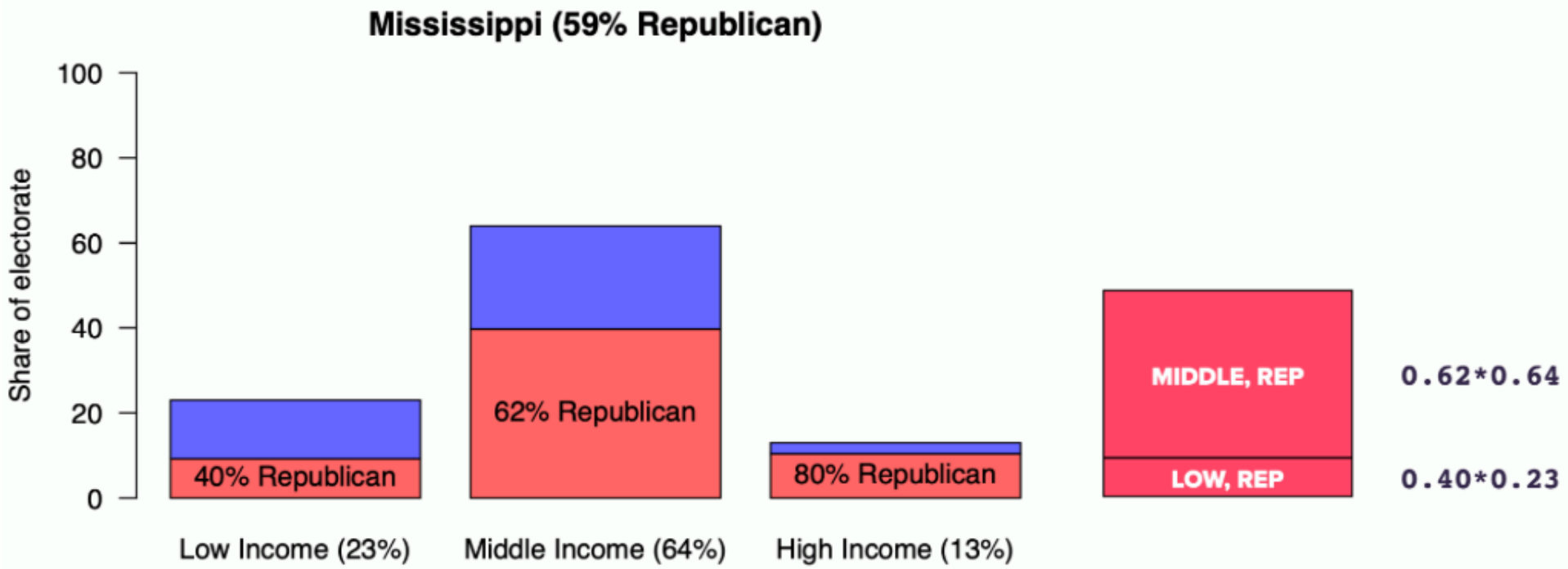
Law of total probability, Mississippi



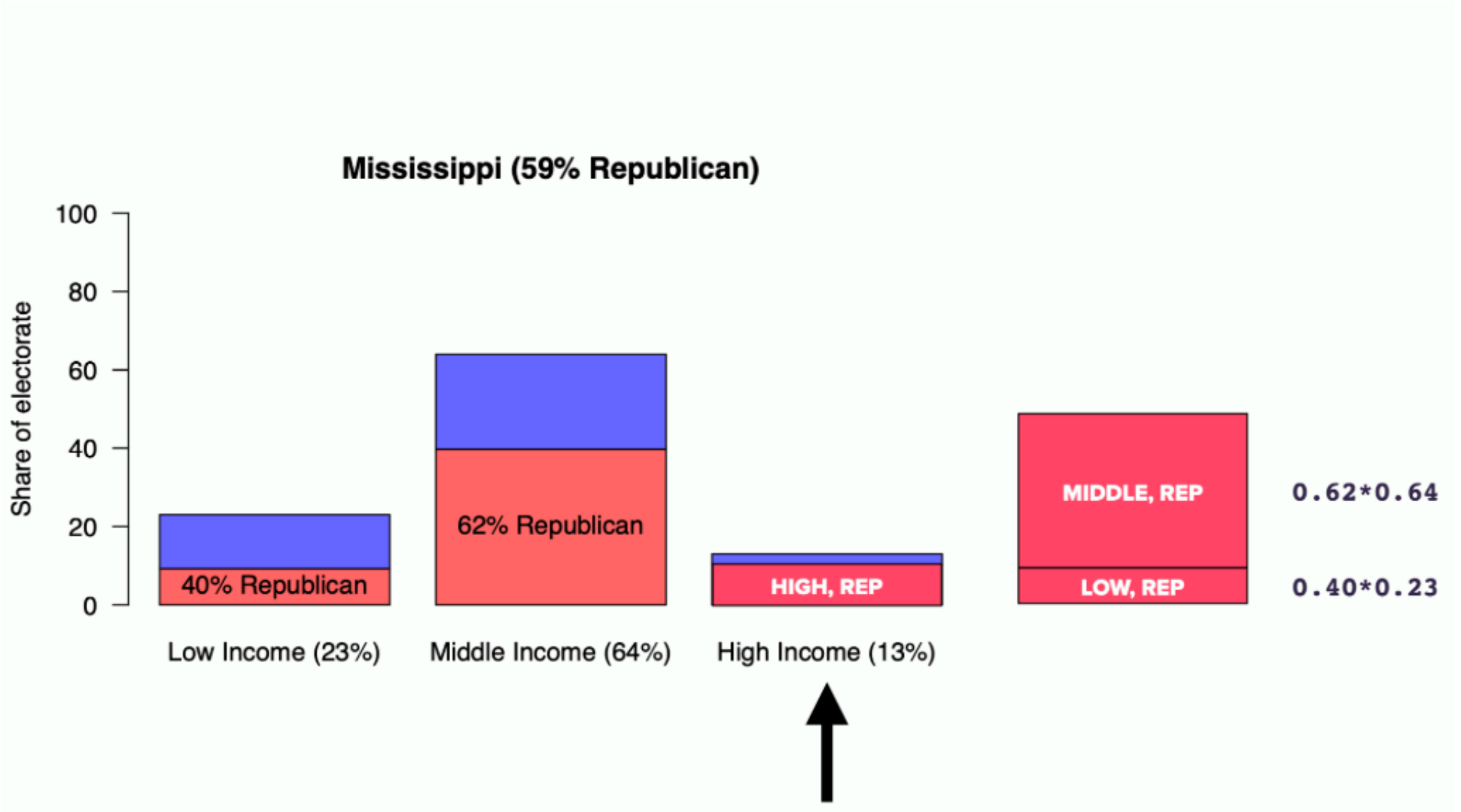
Law of total probability, Mississippi



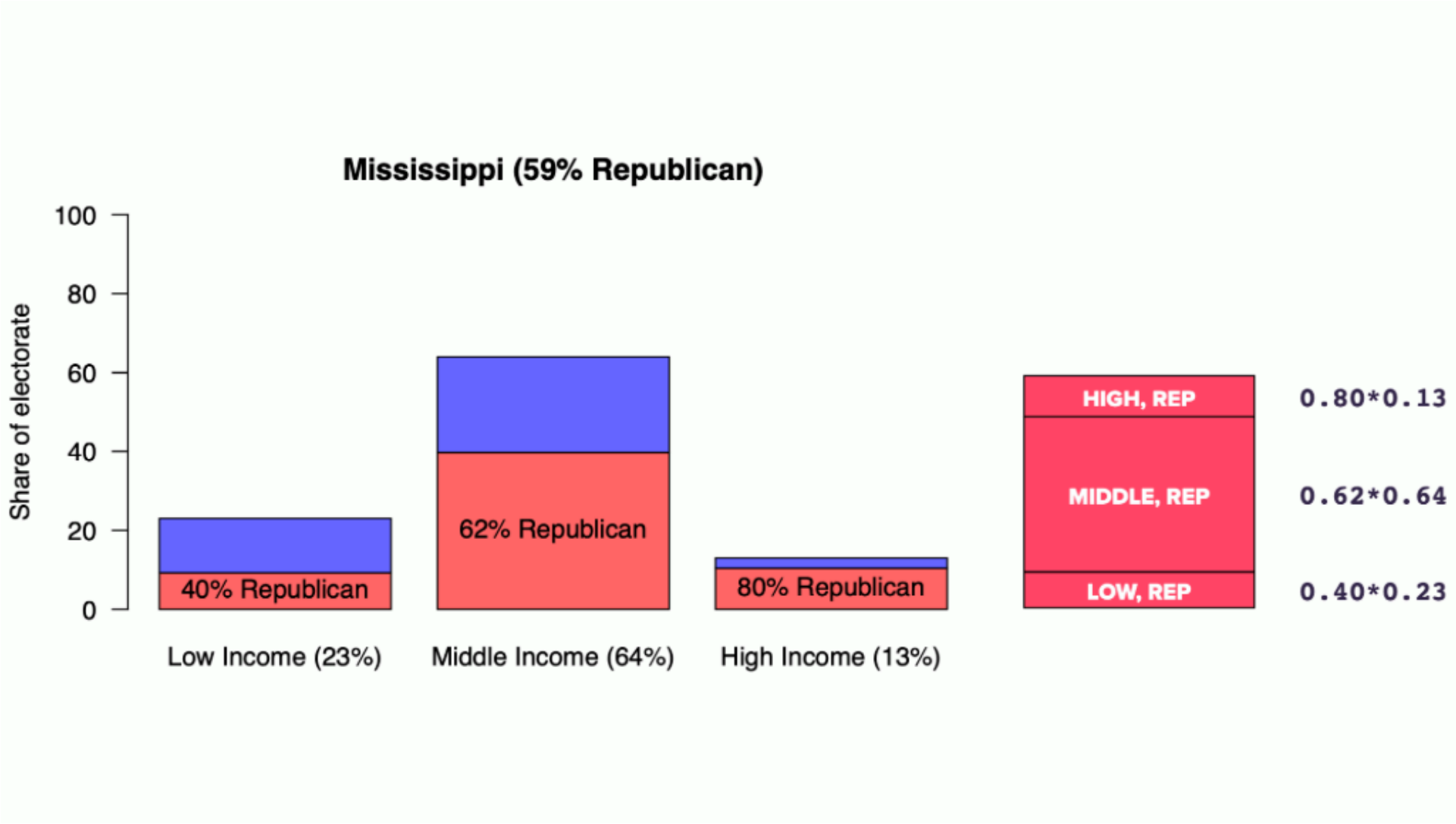
Law of total probability, Mississippi



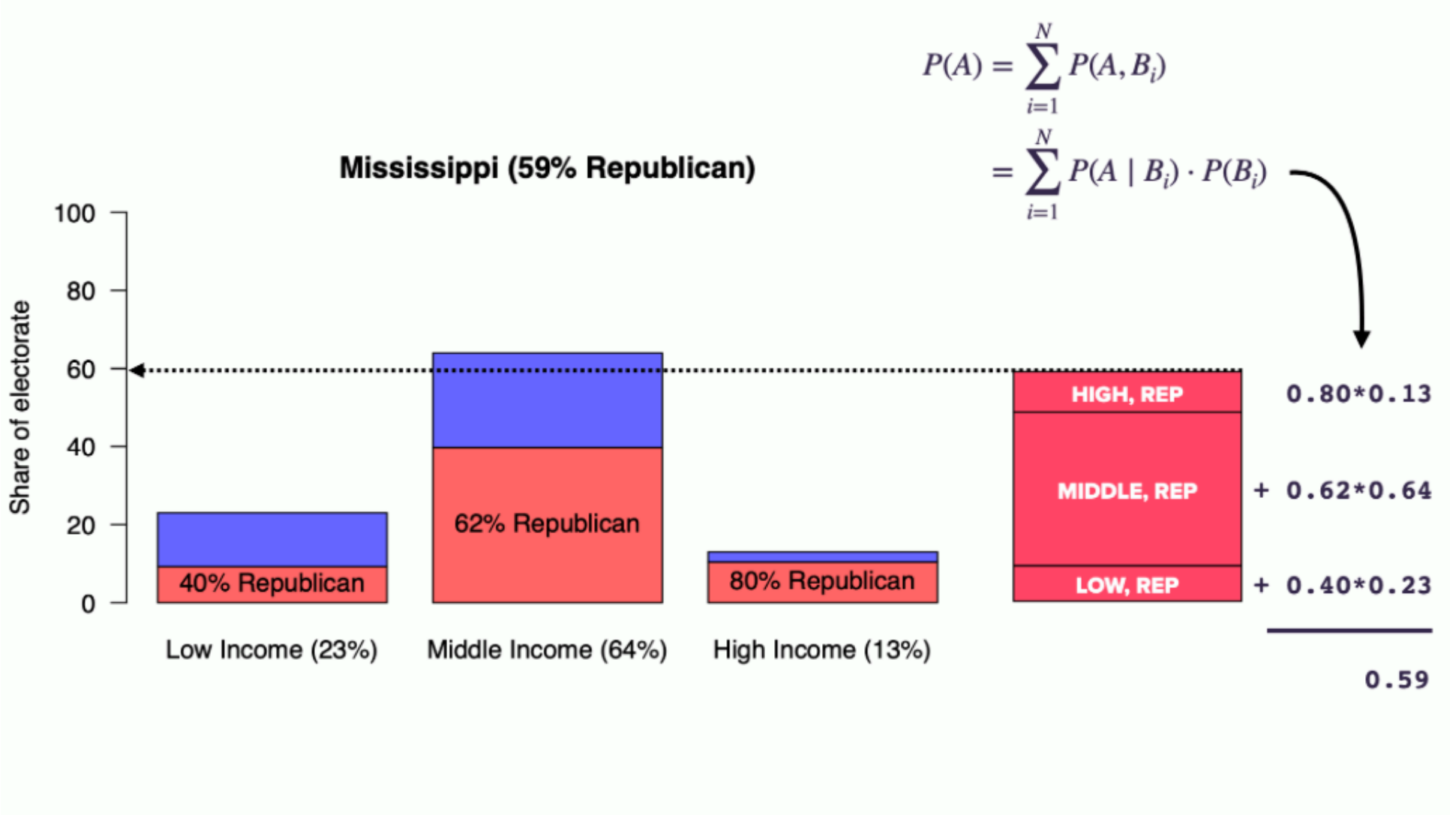
Law of total probability, Mississippi



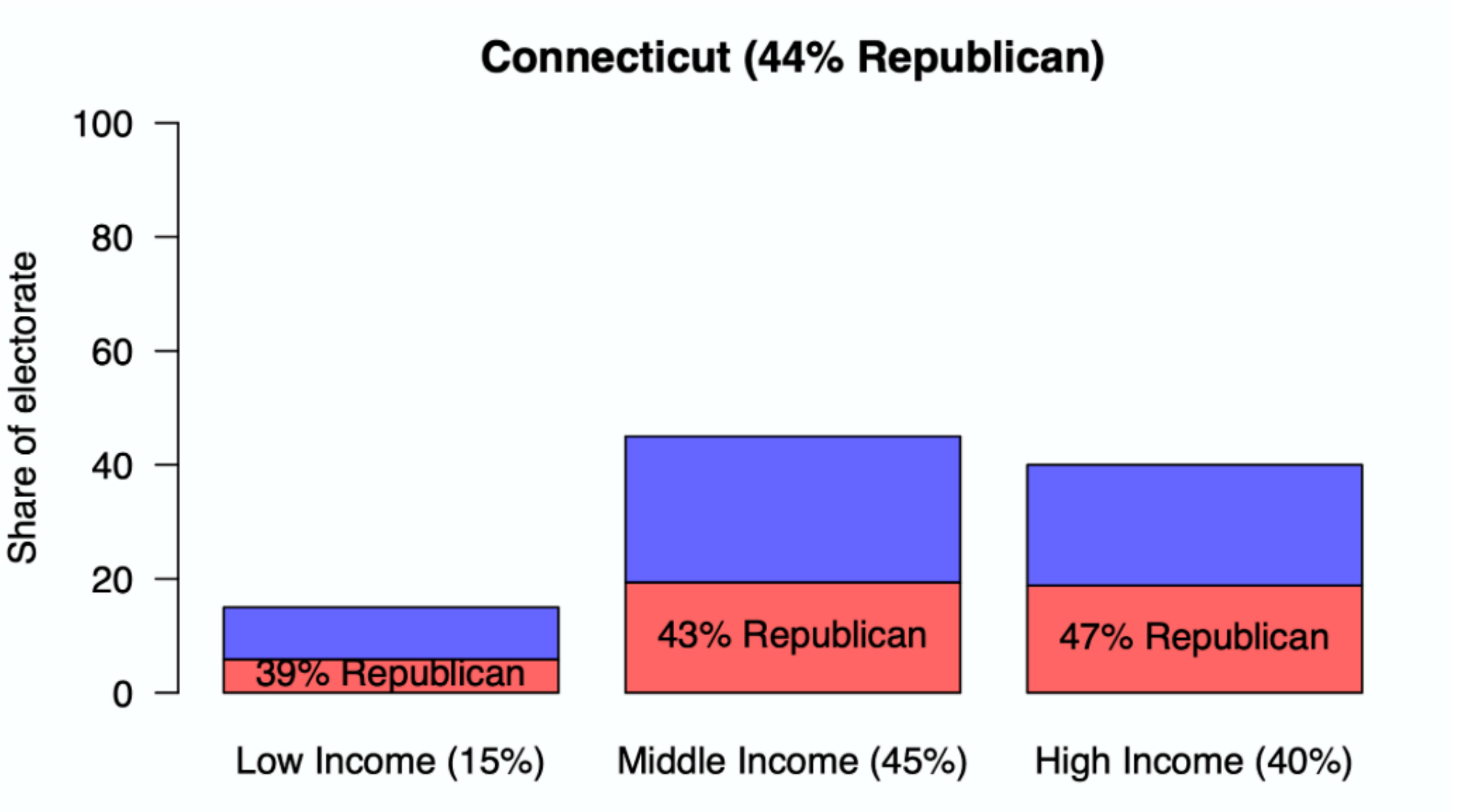
Law of total probability, Mississippi



Law of total probability, Mississippi



And now Connecticut



Connecticut 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:

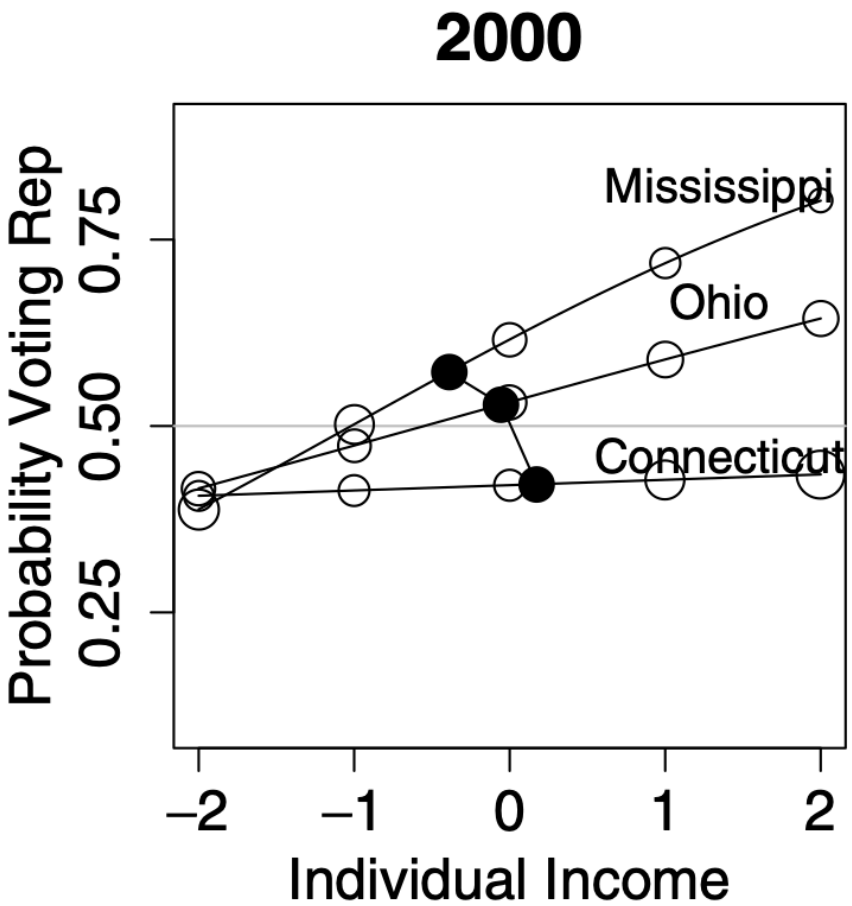
	Low-income	Middle-income	High-income
Connecticut	0.39	0.43	0.47
Mississippi	0.40	0.62	0.80

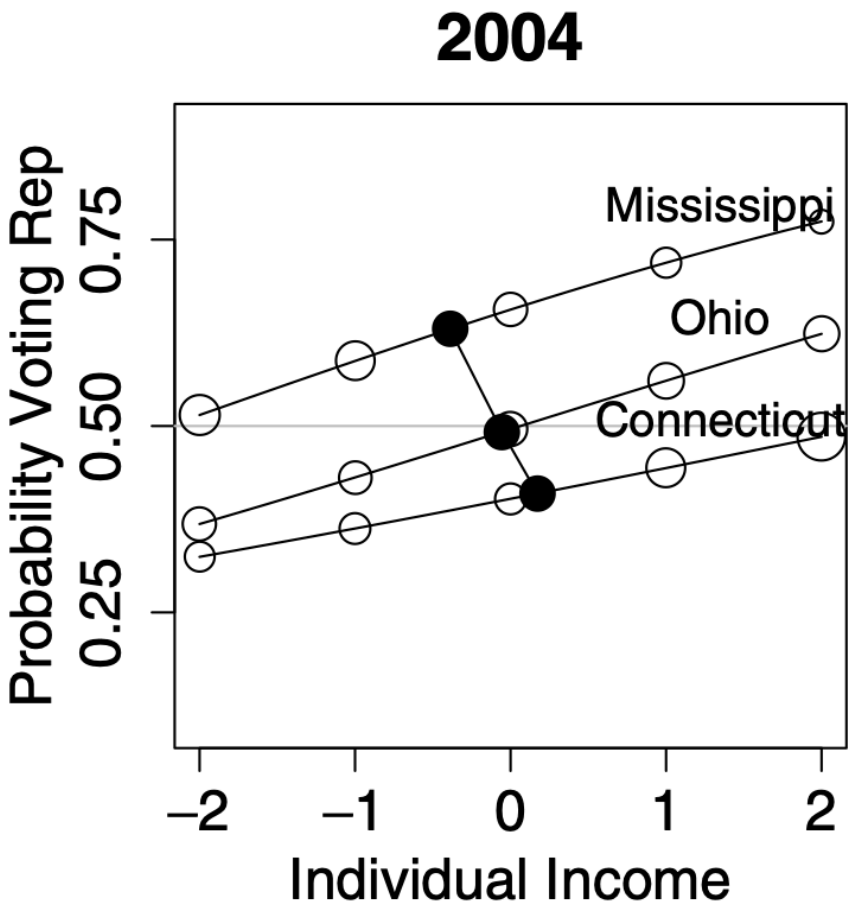
Q: Does income really tell me anything about why CT is blue and MS is red?

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?

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.

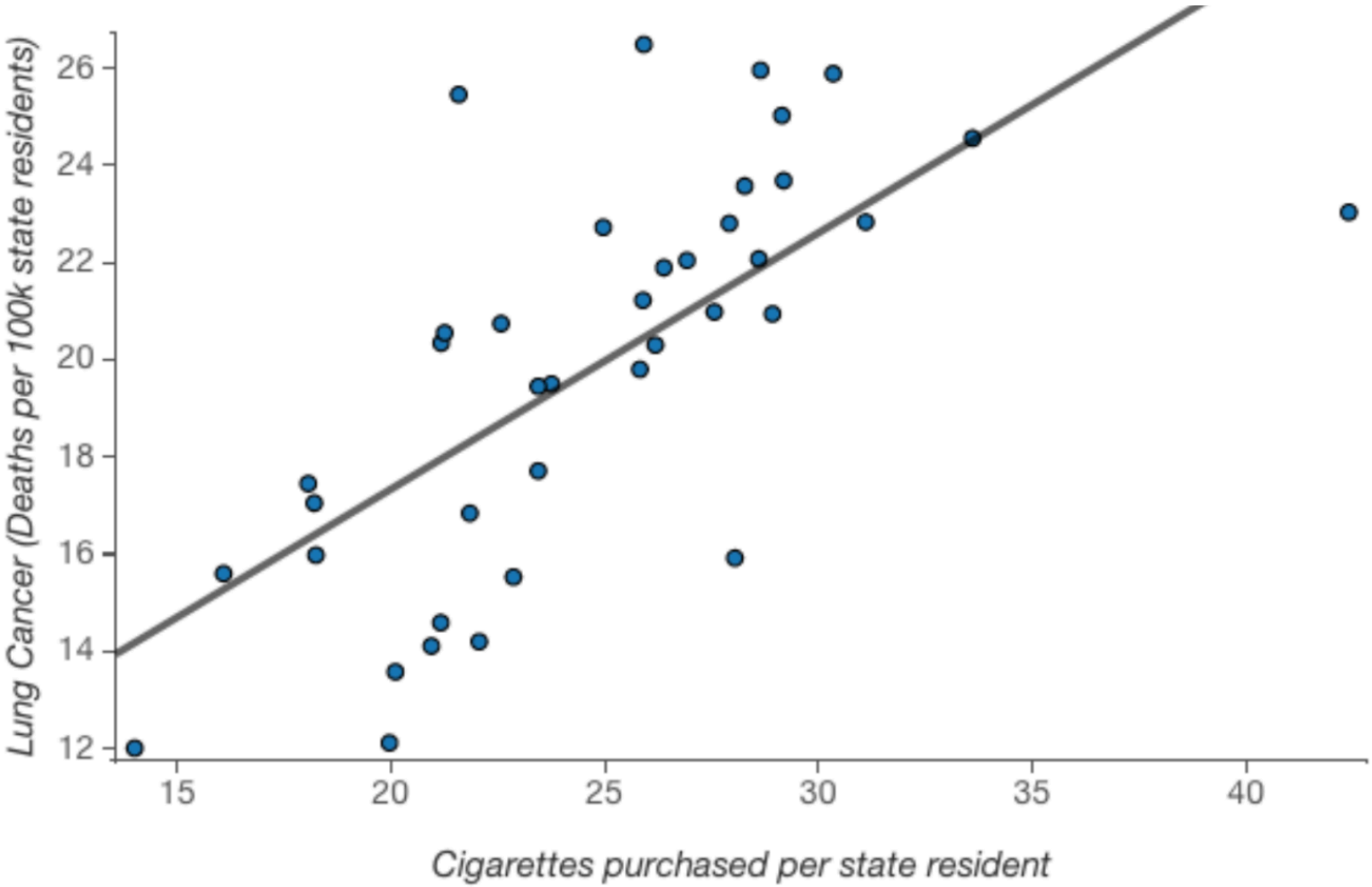
Do groups with higher average levels of A tend to have higher B?

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!!**

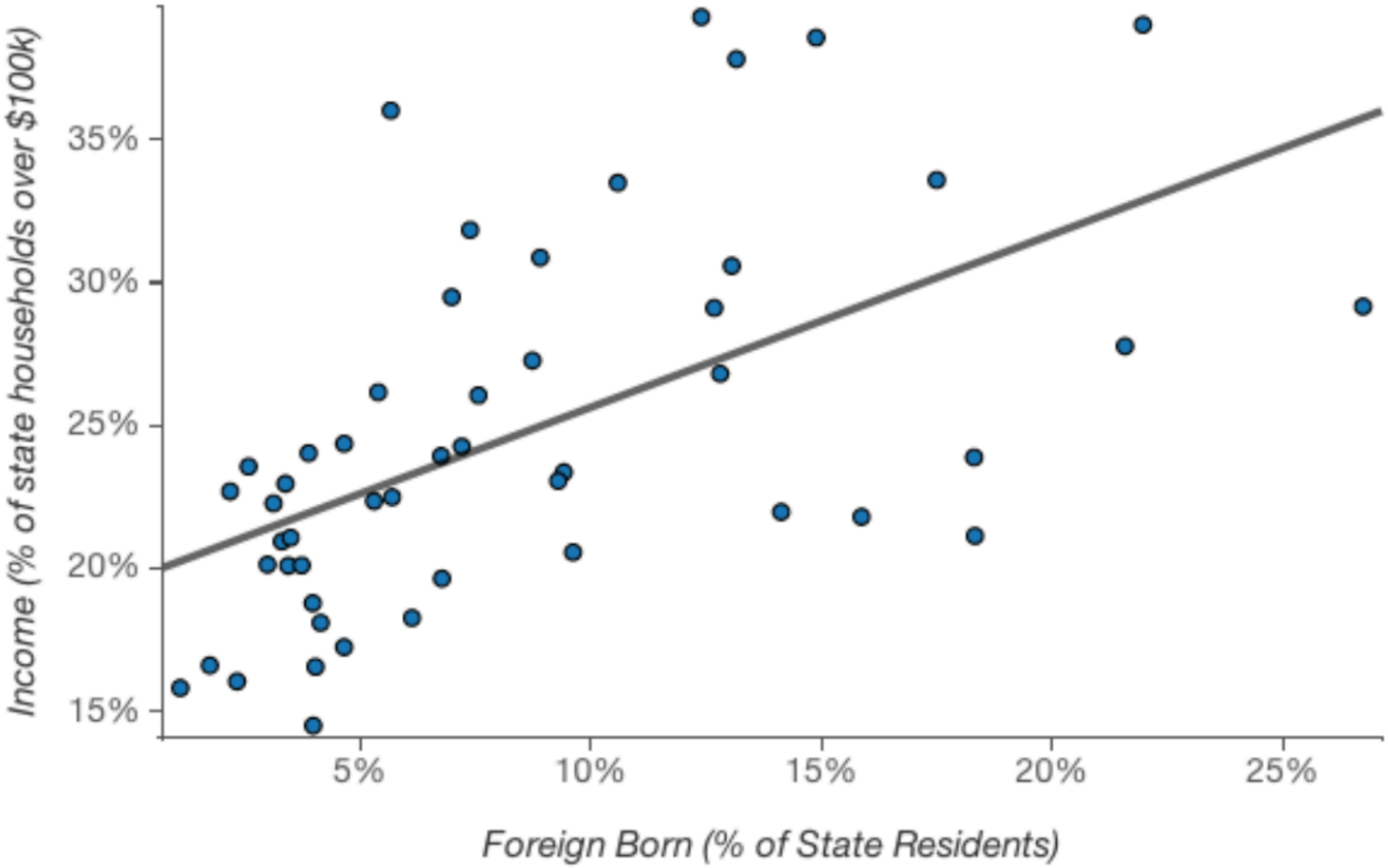
The ecological fallacy

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

Let's play a probability game

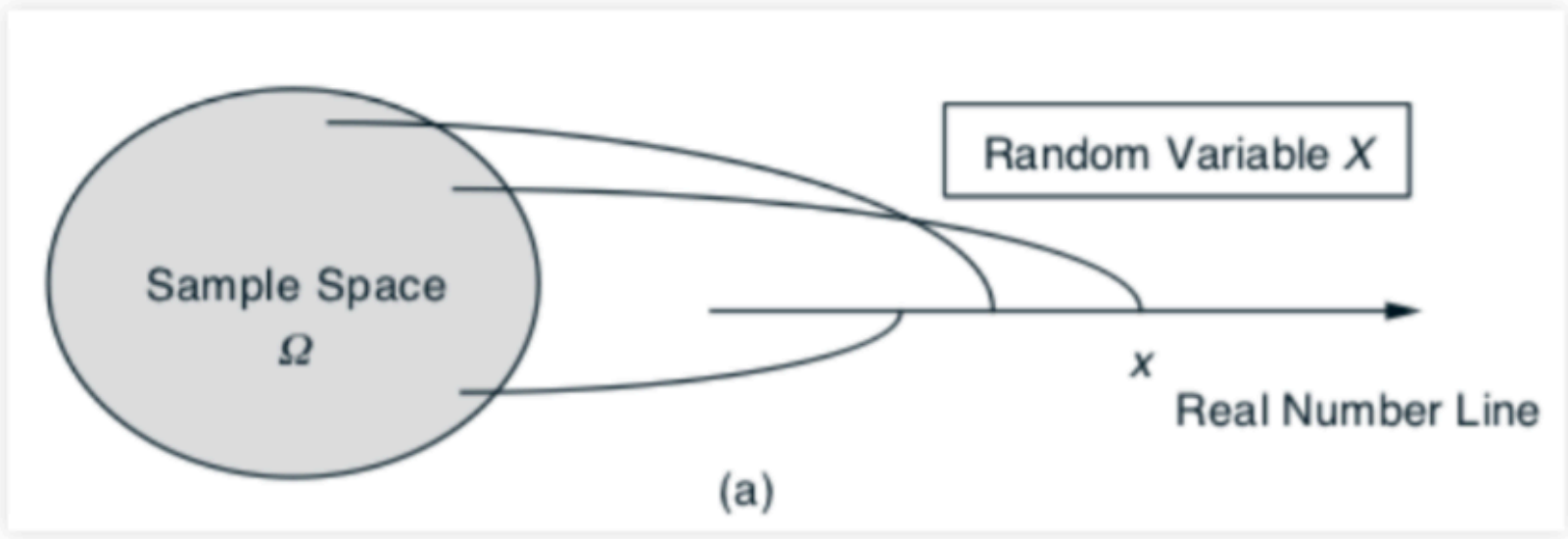
- You will flip a coin ...
- if heads, you will write down the number 1 if your social security number ends with an even digit, otherwise write down 0.
- if tails, you will write down the number 1 if you smoked weed in the past month, otherwise 0.
- Question: What percentage of students have smoked weed in the past month?

Random variables & probability models

From probabilities to random variables

Suppose we have an uncertain outcome with sample space Ω.

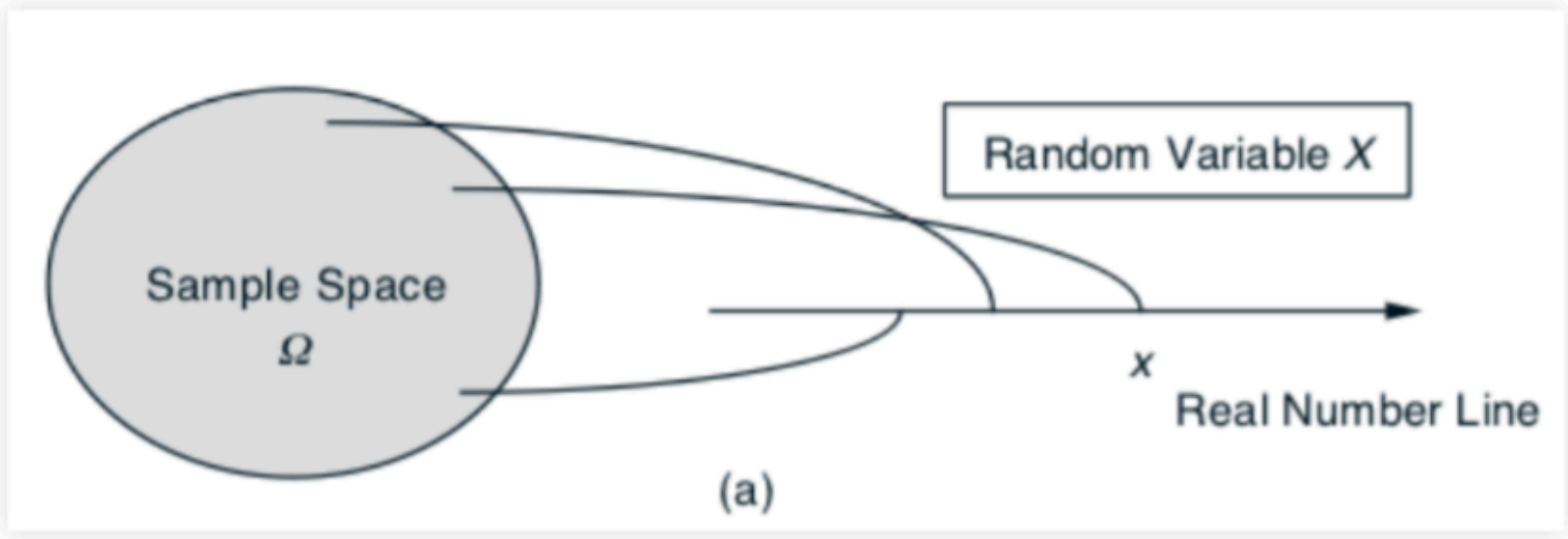
A random variable X is a real-valued function of the uncertain outcome. That is, it maps each element $\omega \in \Omega$ to a real number.



From probabilities to random variables

Suppose we have an uncertain outcome with sample space Ω.

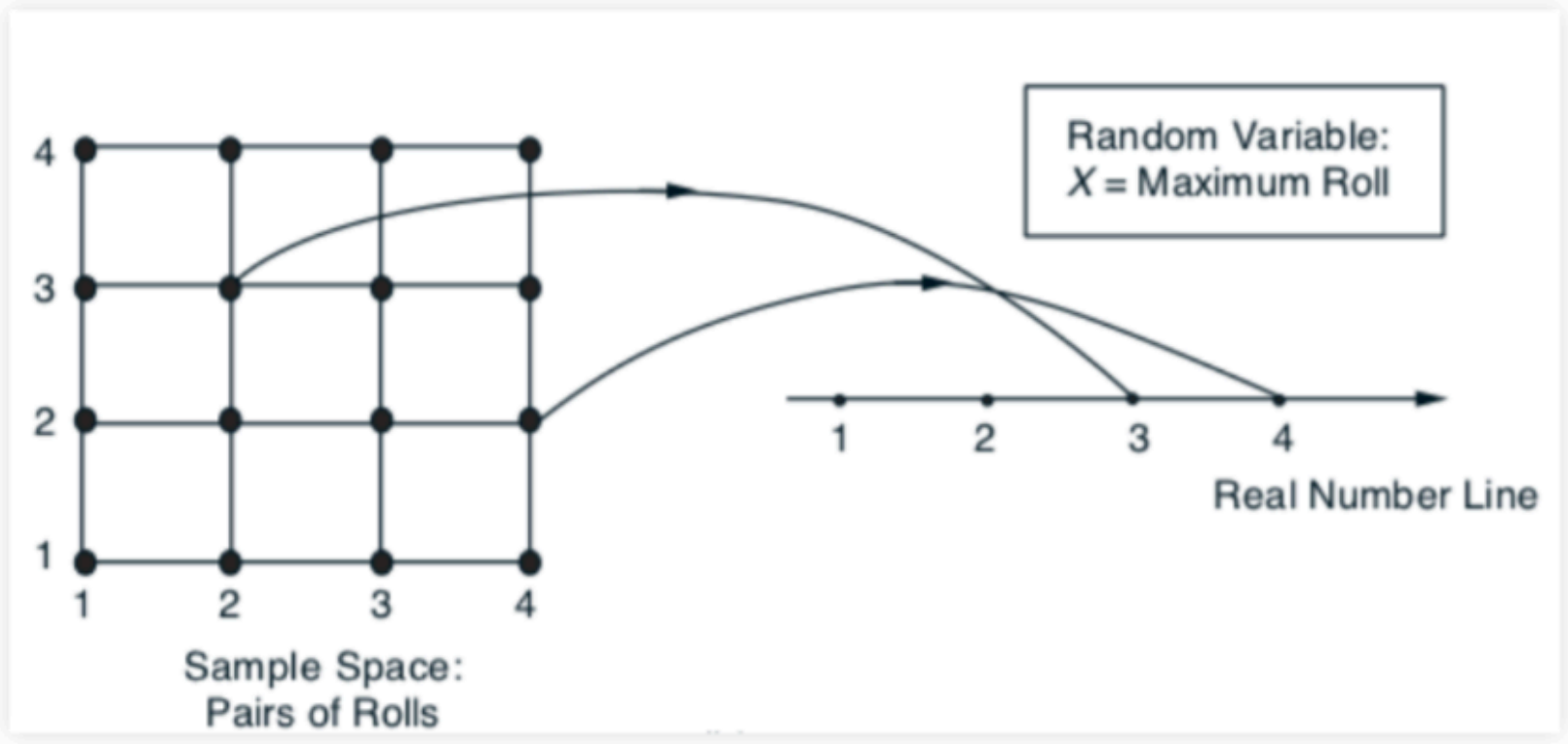
A random variable X is a real-valued function of the uncertain outcome. That is, it maps each element $\omega \in \Omega$ to a real number.



Simple mantra: A random variable is a numerical summary of some uncertain outcome.

Example 1: Rolling two dice

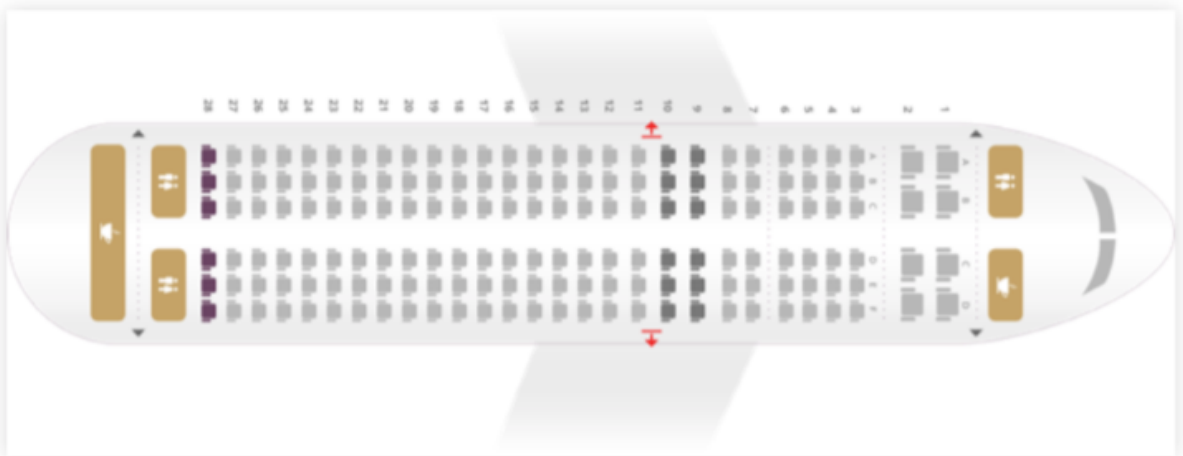
Suppose you roll two dice (an uncertain outcome). An example of a random variable is the maximum of the two rolls:



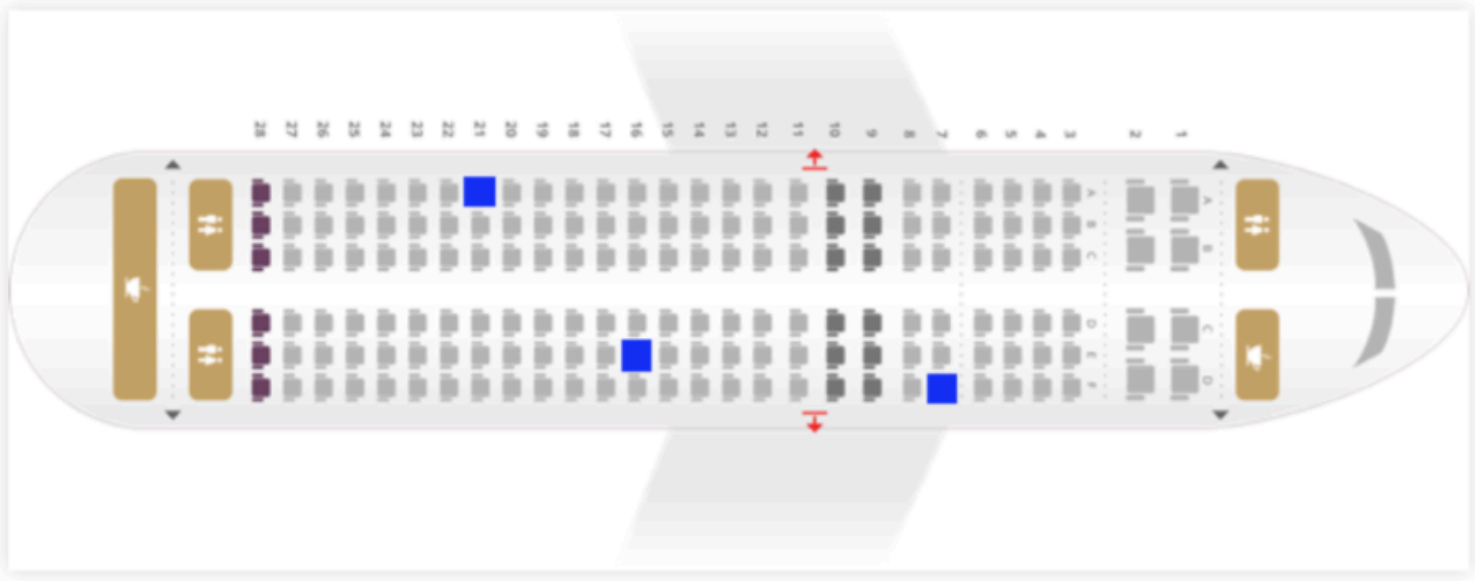
Example 2: Airline no shows

Suppose you're trying to predict the number of no-shows on a flight:

- The sample space Ω is all possible combinations of seats.
- Each $\omega \in \Omega$ is some particular combination of seats that could no-show, e.g. "2A, 13C, 17F"
- The random variable $X(\omega)$ is the size of ω: that is, how many seats no-showed.

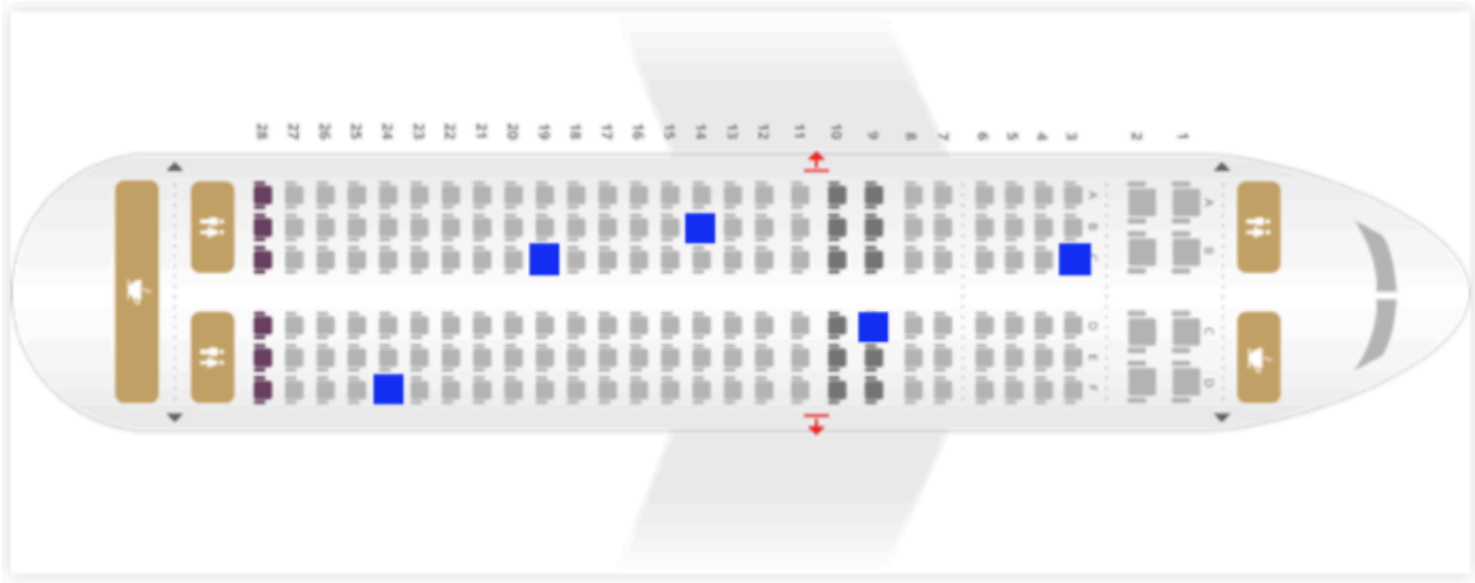


Example 2: Airline no shows



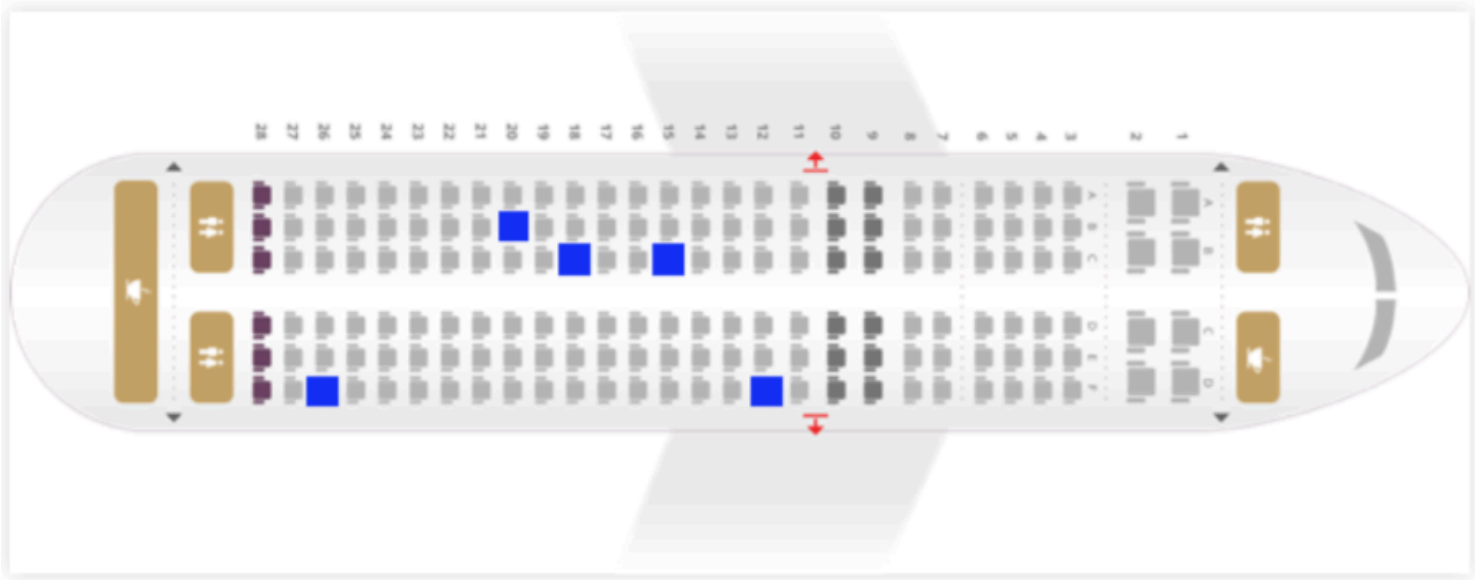
- $\omega =$ 7F, 16E, 21A
- $X(\omega) = 3$

Example 2: Airline no shows



- $\omega = 3C, 9D, 14B, 19C, 24F$
- $X(\omega) = 5$

Example 2: Airline no shows



- $\omega =$ 12F, 15C, 18C, 20B, 26F
- $X(\omega) = 5$

Note: different ω's can map top the same number / summary

Random variables: Big picture summary

- A random variable is a real-valued function (i.e. numerical summary) of the outcome ω. Often this outcome fades into the background, but it's always there lurking.
- We describe the behavior of a random variable in terms of its probability distribution P: a set of possible outcomes for the random variable, together with their probabilities.
- We can associate with each random variable certain "averages" or "moments" of interest (e.g. mean, variance)
- A function of a random variable defines another random variable.

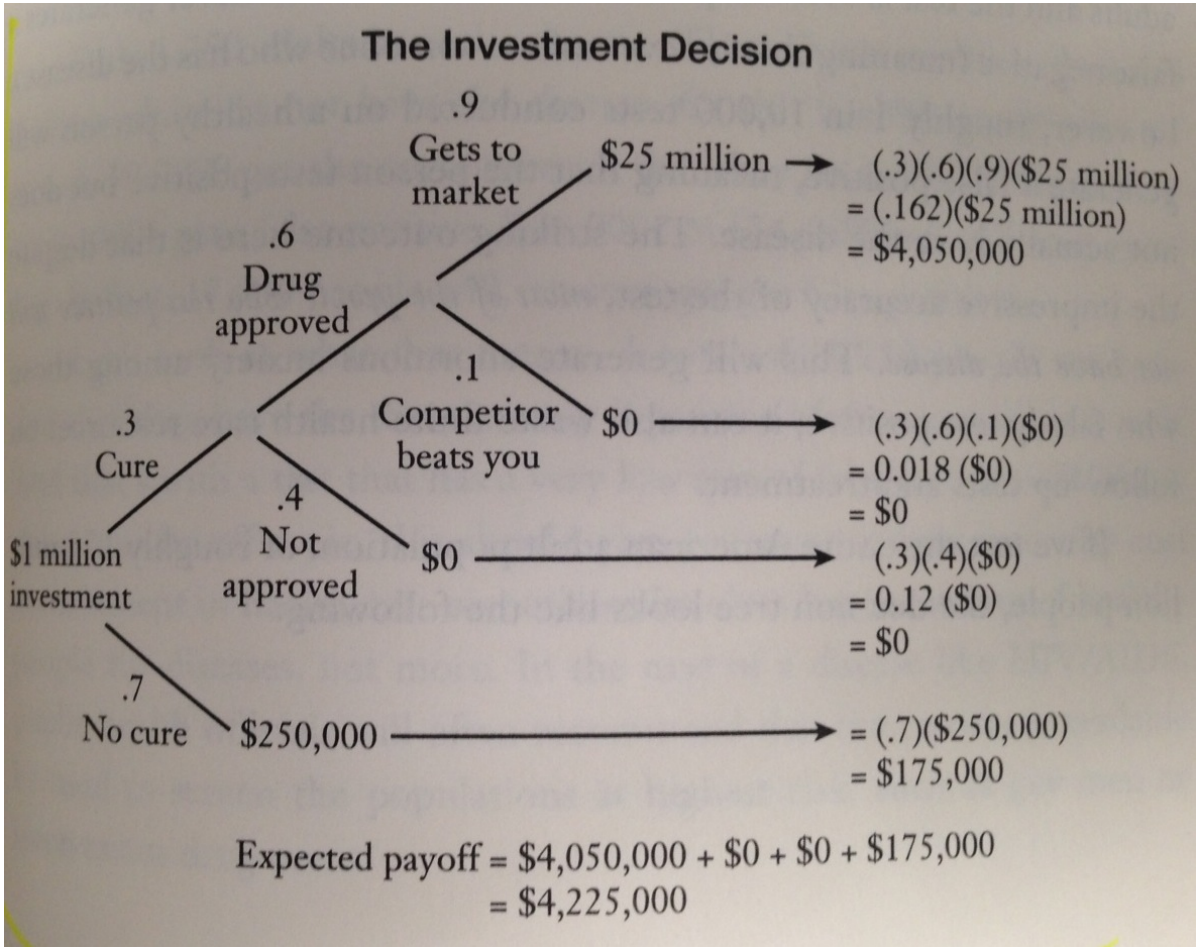
So, now what?

So, now what?

Let's use this framework to make decisions under uncertainty!

Random variables, probability, and decisions

Suppose you are presented with an investment opportunity in the development of a drug... probabilities are a vehicle to help us build scenarios and make decisions.



We have a new **random variable**, i.e, our revenue, with the following probabilities...

Revenue	$P(Revenue)$
\$250,000	0.7
\$0	0.138
\$25,000,000	0.162

The expected revenue is then \$4,225,000...

So, should we invest or not?

What if you could choose this investment instead?

Revenue	$P(Revenue)$
\$3,721,428	0.7
\$0	0.138
\$10,000,000	0.162

The expected revenue is still \$4,225,000...
What is the difference?

Mean and variance of a random variable

The mean or expected value is defined as (for a discrete X):

$$E(X) = \sum_{i=1}^{n} Pr(x_i) \times x_i$$

We weight each possible value by how likely they are... this provides us with a measure of centrality of the distribution... a "good" prediction for X!

Mean and variance of a random variable

The variance is defined as (for a discrete X):

$$Var(X) = \sum_{i=1}^{n} Pr(x_i) \times [x_i - E(X)]^2$$

Weighted average of squared prediction errors... This is a measure of spread of a distribution. More risky distributions have larger variance.

The standard deviation

- What are the units of $E(X)$? What are the units of $Var(X)$?
- A more intuitive way to understand the spread of a distribution is to look at the standard deviation:

$$sd(X) = \sqrt{Var(X)}$$

- What are the units of $sd(X)$?

Targeted marketing

Suppose you are deciding whether or not to target a customer with a promotion (or an advertisement)...

It will cost you \$.80 (eighty cents) to run the promotion and a customer spends \$40 if they respond to the promotion.

Should you do it?

Targeted marketing

Should we send the promotion?

Well, it depends on how likely it is that the customer will respond!

If they respond, you get 40-0.8=$39.20.

If they do not respond, you lose $0.80.

Let's assume your "predictive analytics" team has studied the conditional probability of customer responses given customer characteristics... (say, previous purchase behavior, demographics, etc)

Targeted marketing

Suppose that for a particular customer, the probability of a response is 0.05.

Profits	$P(Profits)$
\$-0.8	0.95
\$39.20	0.05

Should you do the promotion?

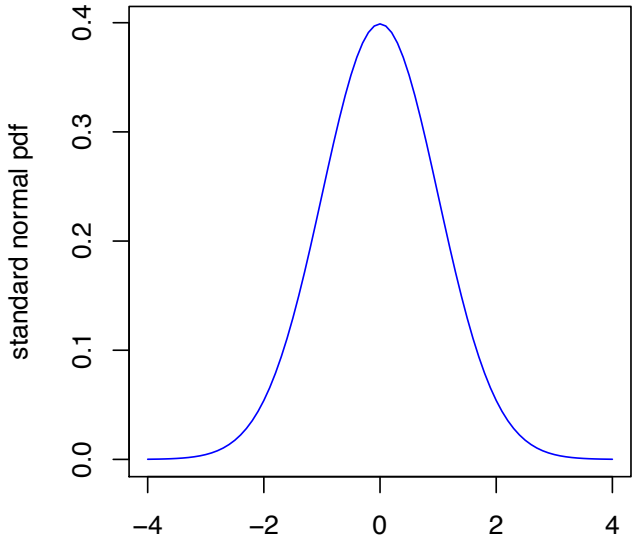
Homework question: How low can the probability of a response be so that it is still a good idea to send out the promotion?

Continuous random variables

- Suppose we are trying to predict tomorrow's return on the S&P500...
- Question: What is the random variable of interest?
- Question: How can we describe our uncertainty about tomorrow's outcome?
- Listing all possible values seems like a crazy task... we'll work with intervals instead.
- These are call continuous random variables.
- The probability of an interval is defined by the area under the **probability density function** aka **pdf**.

The normal distribution

- Recall that a random variable is a number we are NOT sure about but we might have some idea of how to describe (model) its potential outcomes.
- The **normal distribution** is the most widely used probability distribution to describe a random variable.
- The probability that the realized random variable ends up in an interval is given by the area under the curve (**pdf**)

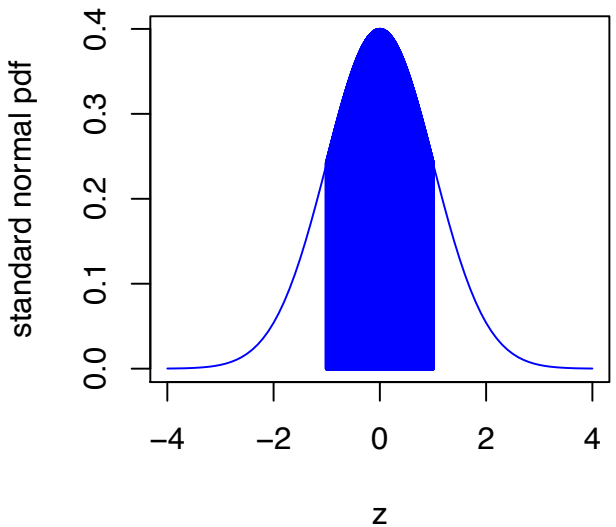


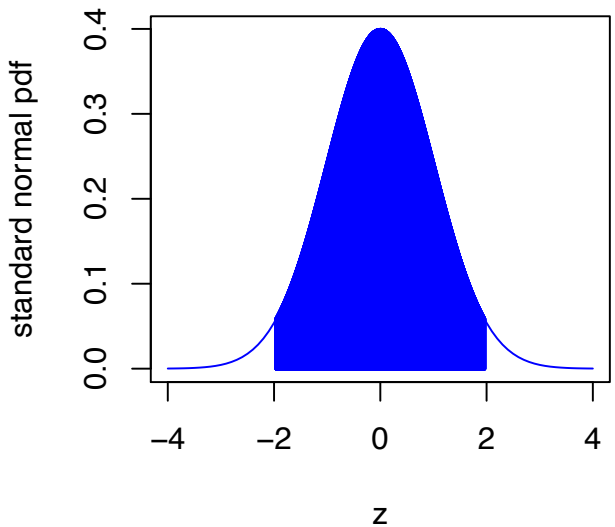
The normal distribution

- The standard normal has mean 0 and variance 1.
- Notation: If $Z \sim N(0,1)$ (Z is the random variable)

$$Pr(-1 < Z < 1) = 0.68$$

$$Pr(-1.96 < Z < 1.96) = 0.95$$





Note:

For simplicity we will often use $P(-2 < Z < 2) \approx 0.95$

Questions:

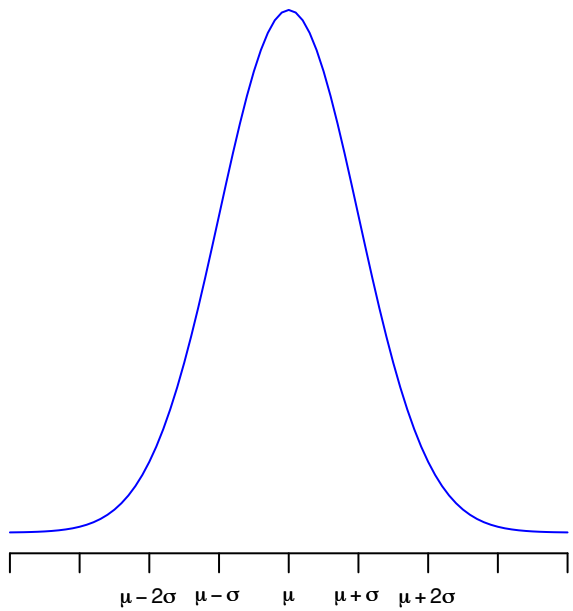
- What is $Pr(Z < 2)$? How about $Pr(Z \leq 2)$?
- What is $Pr(Z < 0)$?

The normal distribution

- The standard normal is not that useful by itself. When we say "the normal distribution", we really mean a family of distributions.
- We obtain pdfs in the normal family by shifting the bell curve around and spreading it out (or tightening it up).

The normal distribution

- We write $X \sim N(\mu, \sigma^2)$. "Normal distribution with mean μ and variance σ^2."
- The parameter μ determines where the curve is. The center of the curve is μ.
- The parameter σ determines how spread out the curve is. The area under the curve in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is 95%. $Pr(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \approx 0.95$

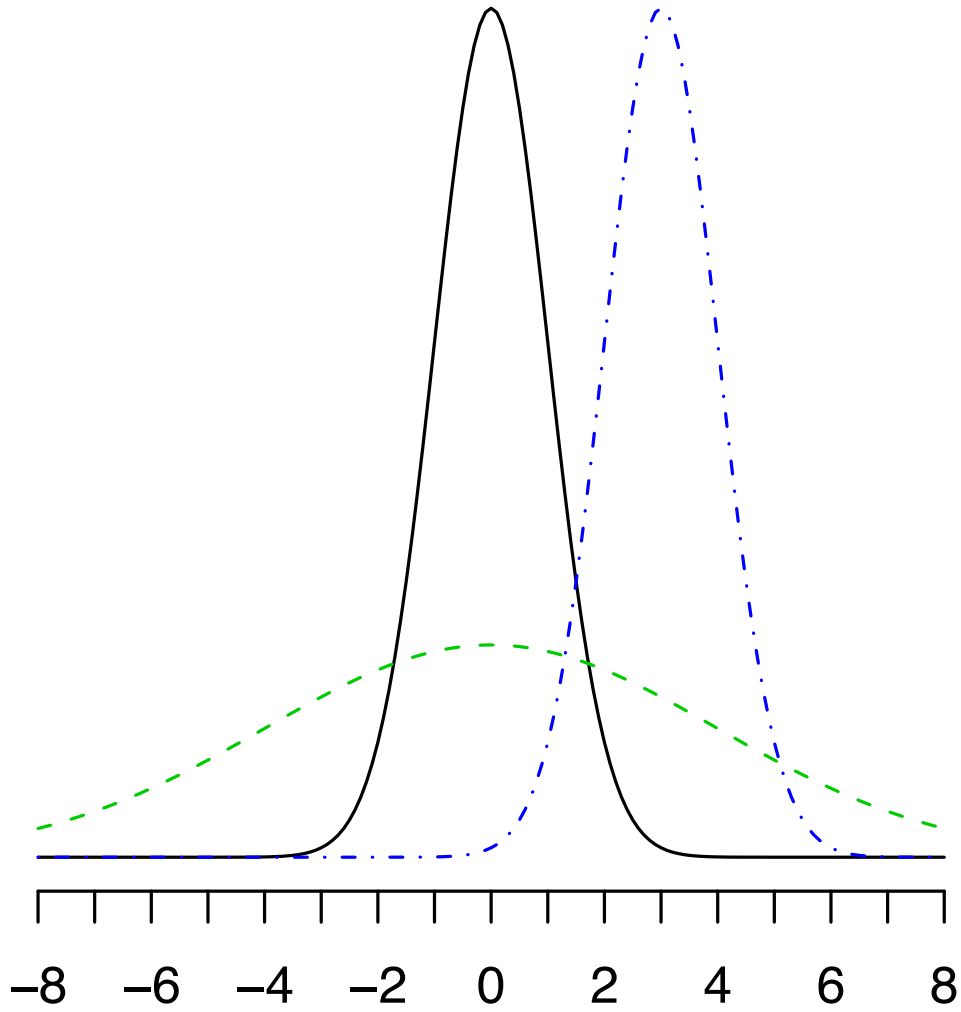


Mean and variance of a random variable

- For the normal family of distributions we can see that the parameter μ talks about "where" the distribution is *located* or *centered*.
- We often use μ as our best guess for a *prediction*.
- The parameter σ talks about how *spread out* the distribution is. This gives us and indication about how *uncertain* or how *risky* our prediction is.
- If X is any random variable, the mean will be a measure of the location of the distribution and the variance will be a measure of how spread out it is.

The normal distribution

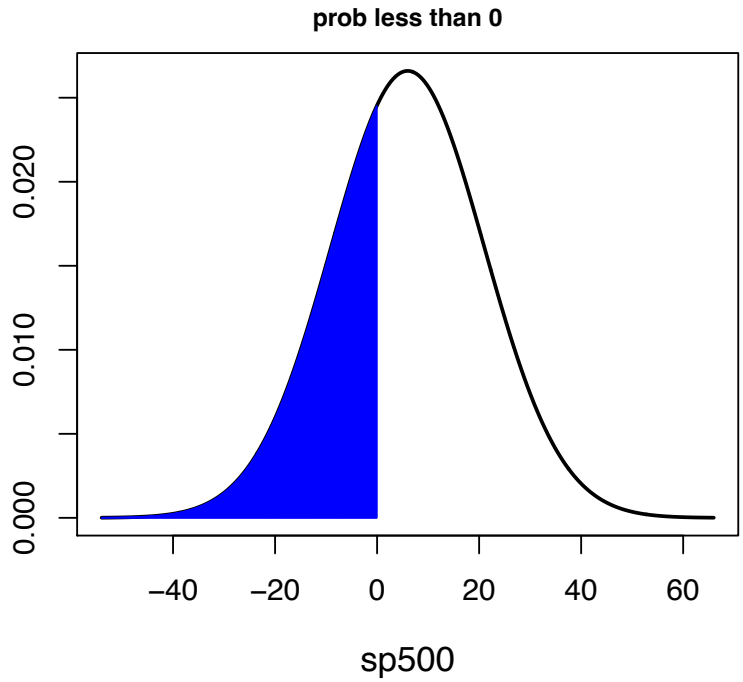
– Example: Below are the pdfs of $X_1 \sim N(0, 1)$, $X_2 \sim N(3, 1)$, and $X_3 \sim N(0, 16)$. Which pdf goes with which X?

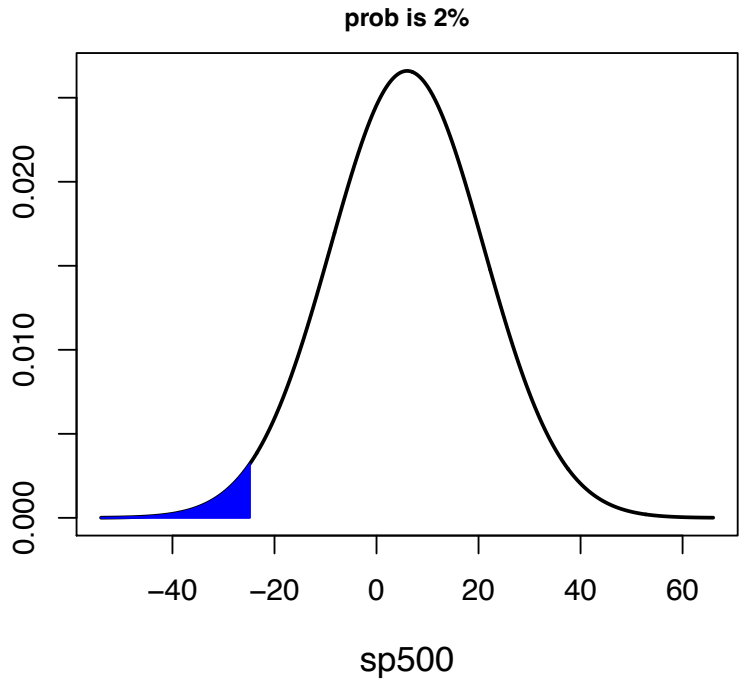


Now back to the S&P500 question

- Assume the annual returns on the SP500 are normally distributed with mean 6% and standard deviation 15%. SP500 $\sim N(6, 225)$. (Notice: $15^2 = 225$).
- Two questions: (i) What is the chance of losing money on a given year? (ii) What is the value that there's only a 2% chance of losing that or more?
- Lloyd Blankfein: *"I spend 98% of my time thinking about 2% probability events!"*
- (i) $Pr(SP500 < 0)$ and (ii) $Pr(SP500 <?) = 0.02$

Now back to the S&P500 question





- (i) $Pr(SP500 < 0) = 0.35$ and (ii) $Pr(SP500 < -25) = 0.02$
- In Excel: NORMDIST and NORMINV (homework!)

The normal distribution – summary

- Note: In
$$X \sim N(\mu, \sigma^2)$$
μ is the mean and σ^2 is the variance.

- Standardization: if $X \sim N(\mu, \sigma^2)$ then
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Summary:
$X \sim N(\mu, \sigma^2)$:
μ: where the curve is
σ: how spread out the curve is
95% chance $X \in \mu \pm 2\sigma$.

The normal distribution – another example

Prior to the 1987 crash, monthly S&P500 returns (r) followed (approximately) a normal with mean 0.012 and standard deviation equal to 0.043. How extreme was the crash of -0.2176? The standardization helps us interpret these numbers...

$$r \sim N(0.012, 0.043^2)$$

$$z = \frac{r - 0.012}{0.043} \sim N(0, 1)$$

For the crash,

$$z = \frac{-0.2176 - 0.012}{0.043} = -5.27$$

How extreme is this $z$value? 5 standard deviations away!!

Regression to the mean

- Imagine your performance on a task follows a standard normal distribution, i.e., $N(0, 1)$... Say you perform that task today and score 2.
- If you perform the same task tomorrow, what is the probability you are going to do worse? 97.5%, right?
- This is called regression to the mean!

Make sure to read the article on this topic available in the class website...