TEXAS McCombs

MASTER OF SCIENCE IN BUSINESS ANALYTICS

Probability

Outline

The basics and conditional probability

Independence

Paradoxes, mixtures, and the rule of total probability

Random variables, distributions, and simulation

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

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:

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

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*?
- 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 Bijan Robinson (@bijan_robinson)?

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 and probability models

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 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

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

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":

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

Some set theory concepts

We need some basic set-theory concepts to make sense of probability, since the sample space Ω is a set, and since "events" are subsets of Ω.

Union: $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$

Intersection: $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$

Complement: $A^C = \tilde{A} = \{\omega : \omega \notin A\}$

Difference: $A \setminus B = \{\omega : \omega \in A, \omega \notin B\}$

Disjointness: A and B are disjoint if $A \cap B = \emptyset$ (the empty set).

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?

Calculation

Now that we have an understanding of the axioms, notation, and interpretation, how do we 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.

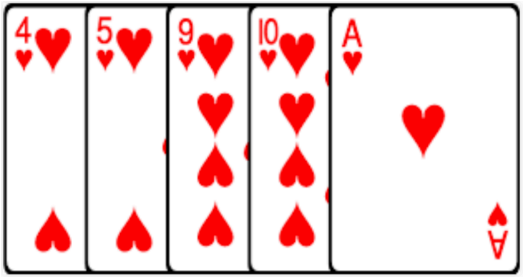
Then for each set $A \subset \Omega$,

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

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.

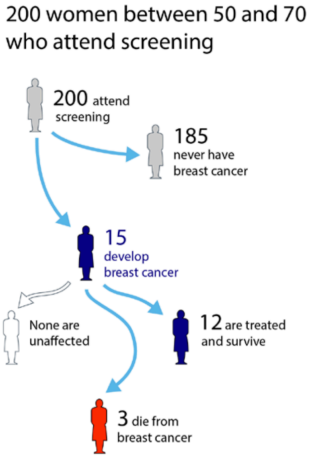
– How many possible flushes are there? Let's start with hearts: $\rightarrow$ There are 13 hearts
$\rightarrow$ To make a flush with hearts, you need any 5 of these 13 cards.
$\rightarrow$ Thus there are $\binom{13}{5}$ = 1287 possible flushes with hearts.
$\rightarrow$ The same argument works for all 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?

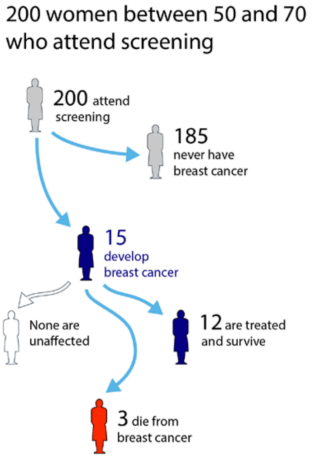
Probability trees are very useful for this task! This involves counting at different levels of the tree.

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:

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

This is actually a new axiom

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

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

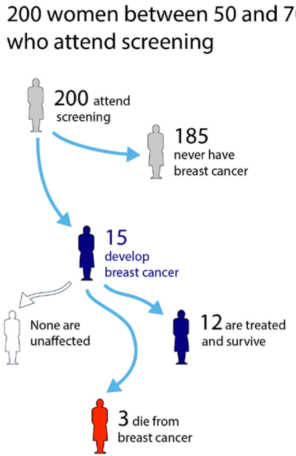
Alternate version

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)

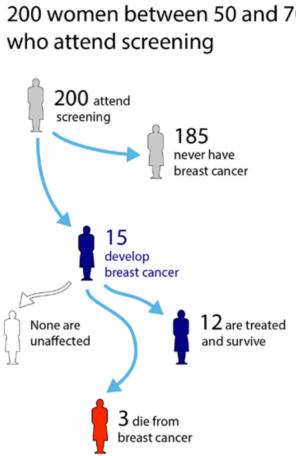


- $P(\text{cancer}) = \frac{15}{200}$
- $P(\text{die,cancer}) = \frac{3}{200}$
- $P(\text{die} \mid \text{cancer}) = \frac{3}{15}$

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

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

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**?

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

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

- What is unknown (A): Loren likes Saving Private Ryan
- What is known (B): Loren 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. Jake Francis	Yes	Yes
2. Alex Stein	No	Yes
3. Grace Deng	Yes	No
4. Jeremy Chen	No	No
5. Charles Waffle	Yes	No
6. Jeremy Ward	Yes	Yes
⋮	⋮	⋮
1575. Michelle Vu	No	Yes
1576. JJ Robinson	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

To figure out a new user's preferences, given that they've watched BoB, compute the following conditional probability:

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

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:

- $P(\text{flip heads second time} \mid \text{flip heads first time}) = P(\text{flip heads second time})$
- $P(\text{stock market up} \mid \text{bird poops on your car}) = P(\text{stock market up})$
- $P(\text{God exists} \mid \text{Longhorns win title}) = P(\text{God exists})$

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.

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.

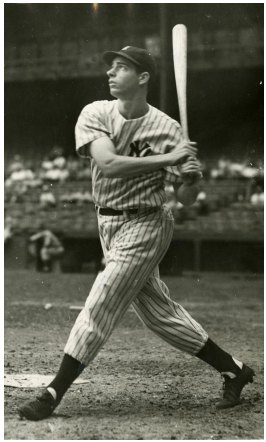
A = Alice has blue eyes.

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 $\iff$ 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.

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?**

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 ...

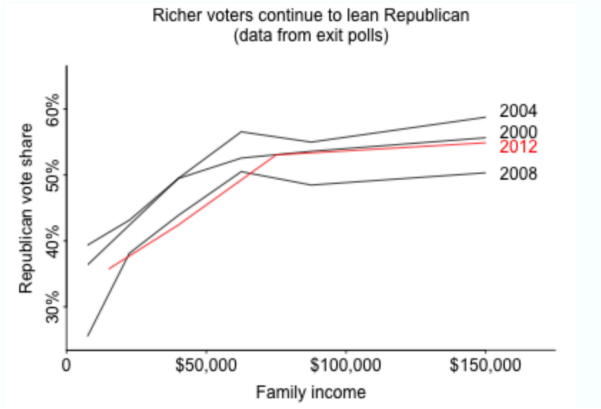
The second paradox

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:
  → 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.

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

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

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

$P(B_i, B_j) = 0 \quad \forall i \neq j \quad$ 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:

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

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:

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!

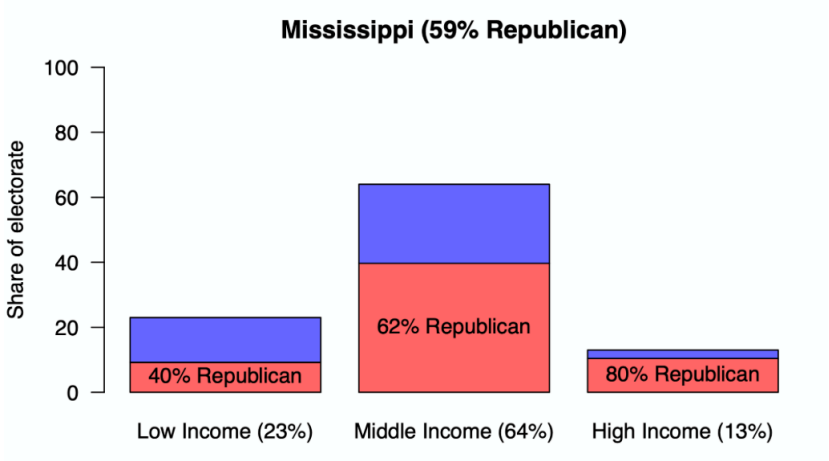
Moral of the story:

- 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.

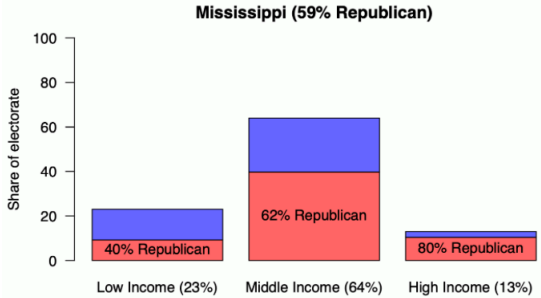
Back to the second paradox

- 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?

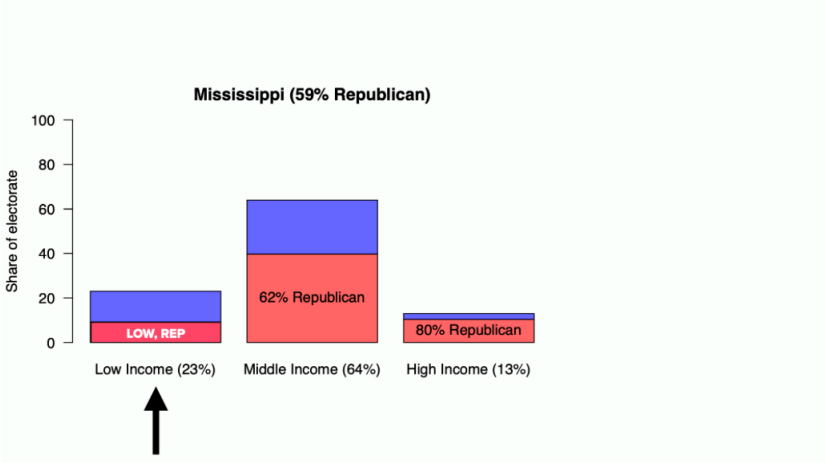
Law of total probability, Mississippi



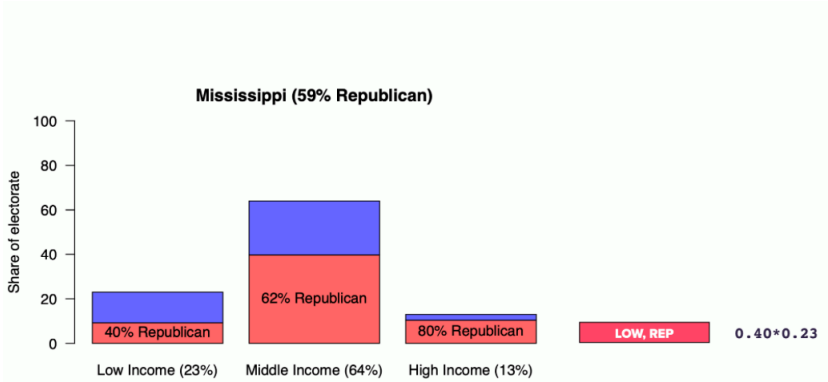
Law of total probability, Mississippi



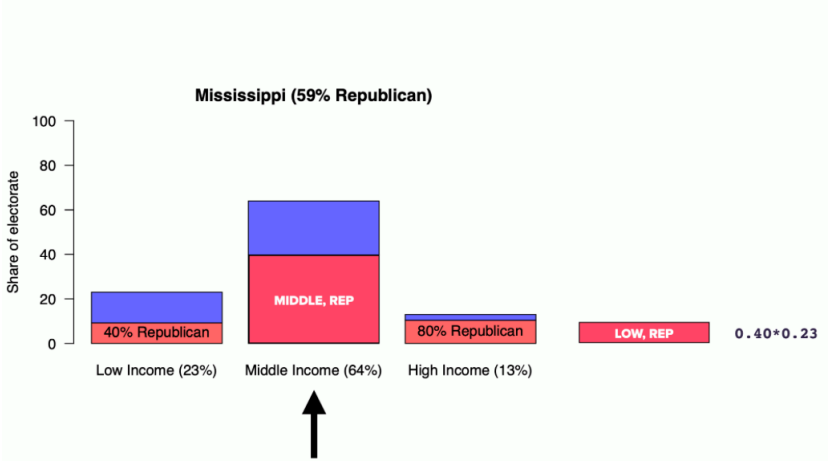
Law of total probability, Mississippi



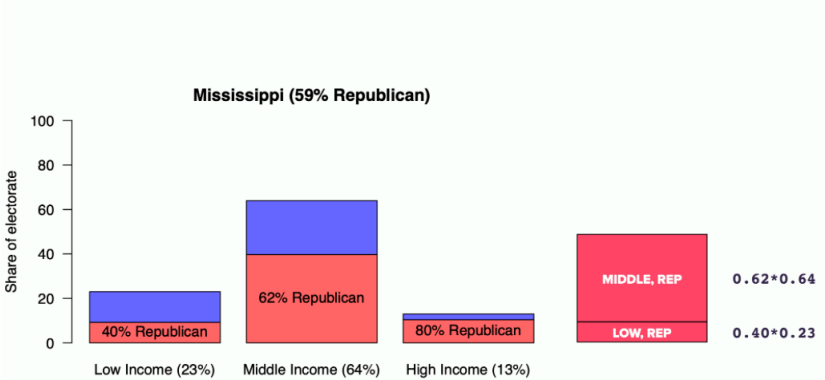
Law of total probability, Mississippi



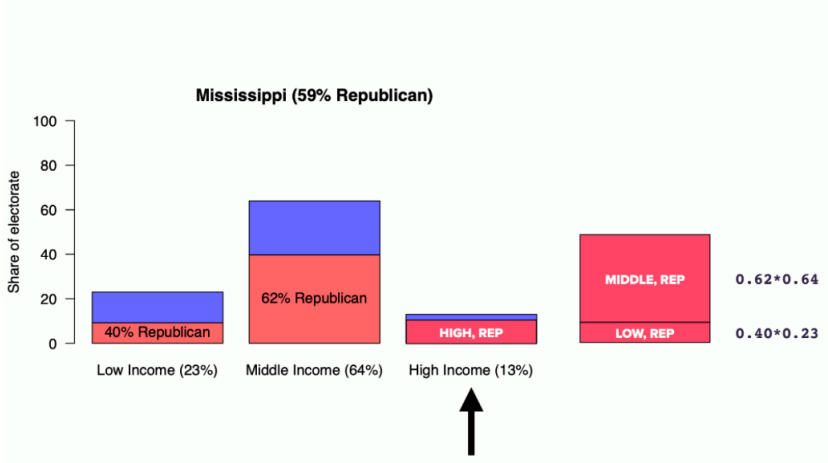
Law of total probability, Mississippi



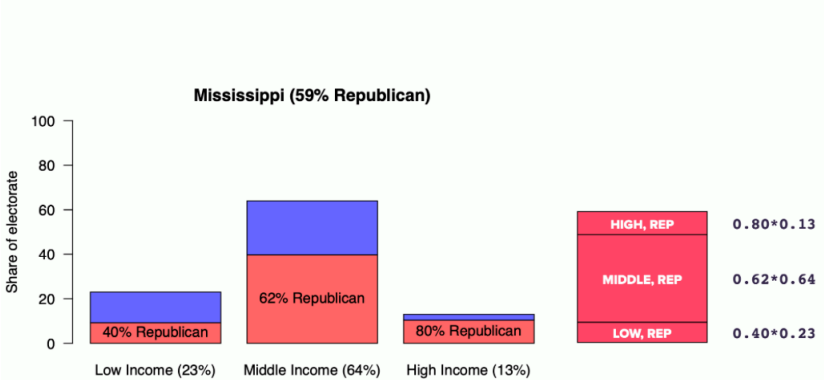
Law of total probability, Mississippi



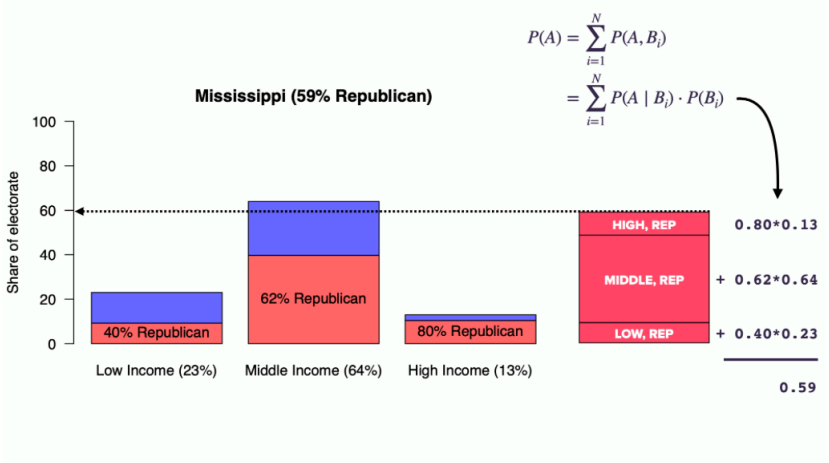
Law of total probability, Mississippi



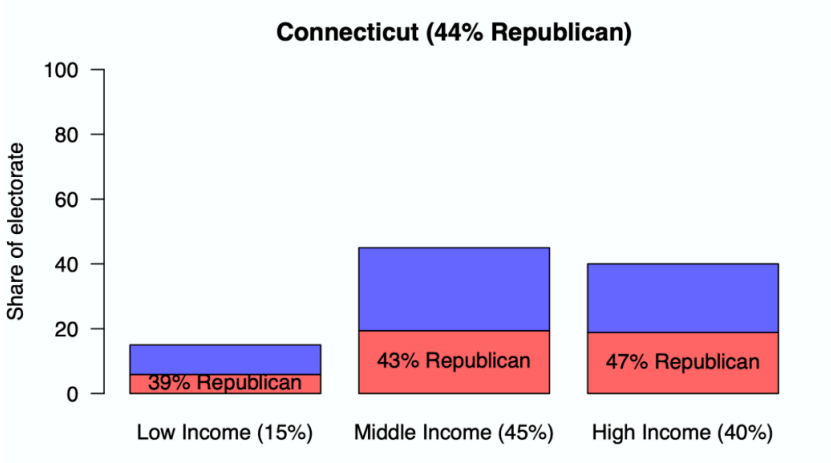
Law of total probability, Mississippi



Law of total probability, Mississippi



And now Connecticut



Connecticut and Mississippi

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

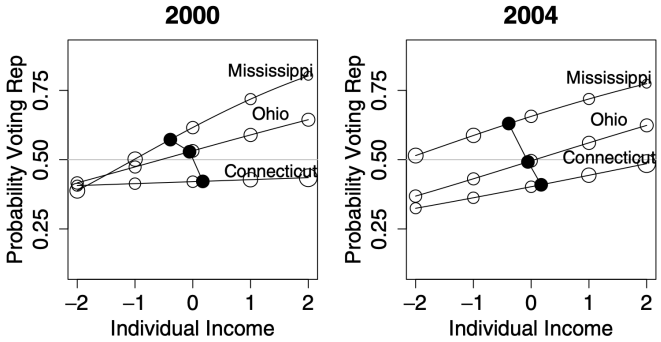
	Low-income	Middle-income	High-income
Connecticut	0.39	0.43	0.47
Mississippi	0.40	0.62	0.80

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.

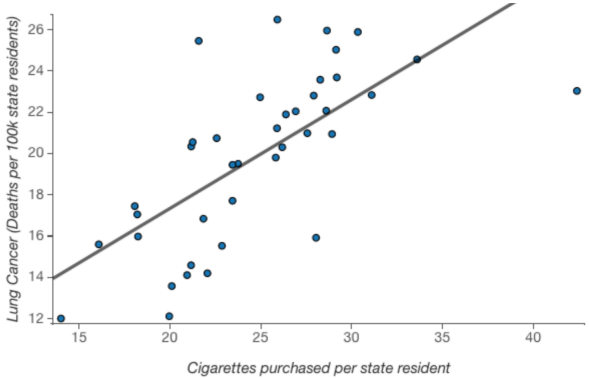
Do groups with higher average levels of A tend to have higher B?

The ecological fallacy: 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. $\leftarrow$ **not necessarily!!**

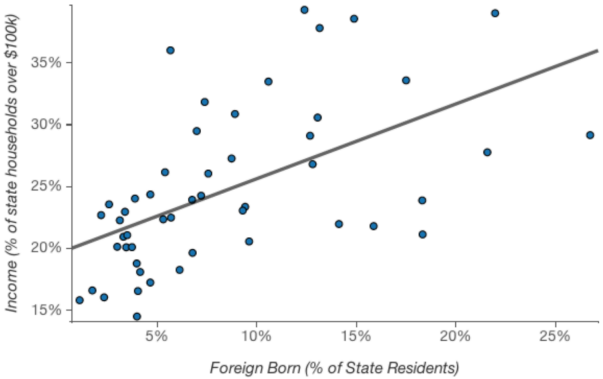
The ecological fallacy

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*.

- 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?

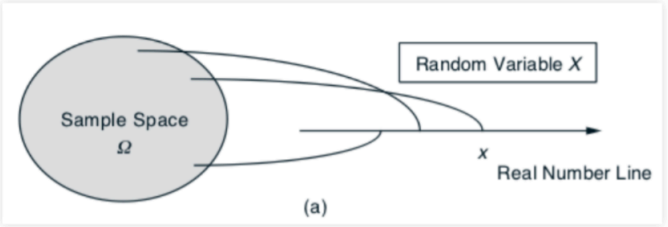
  → Avoid the ecological fallacy: learn to be skeptical when group-level trends are applied to individuals

Random variables, distributions, and simulation

From probabilities to random variables

Suppose we have an uncertain outcome with sample space Ω.

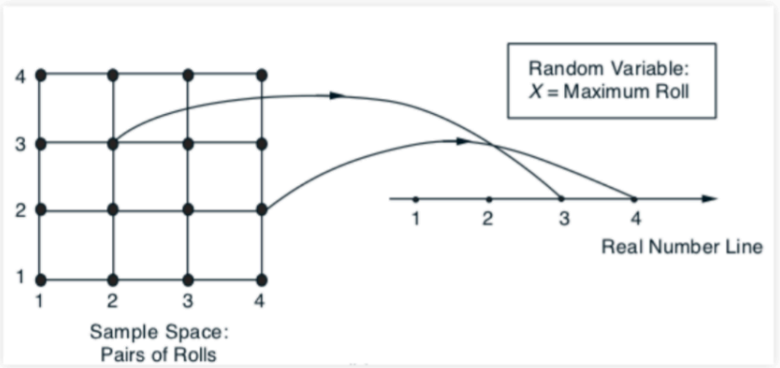
A random variable X is a real-valued function of the uncertain outcome. That is, it maps each element $\omega \in \Omega$ to a real number.



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

Example 1: Rolling two dice

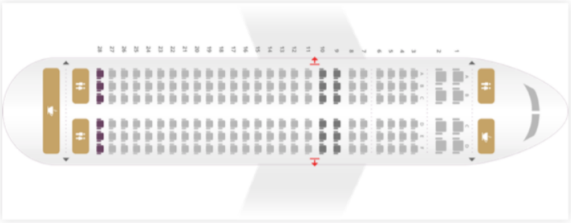
Suppose you roll two dice (an uncertain outcome). An example of a random variable is the maximum of the two rolls:



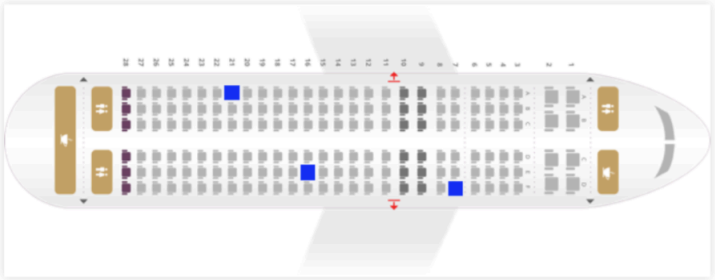
Example 2: Airline no shows

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.

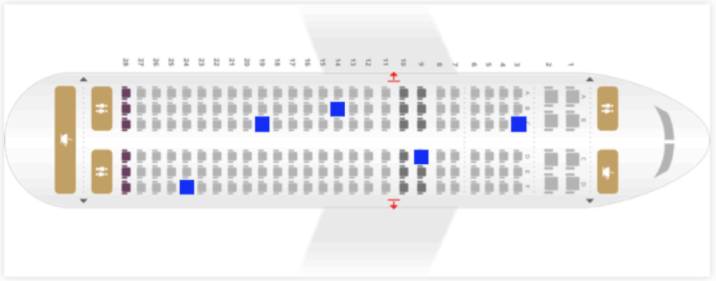


Example 2: Airline no shows



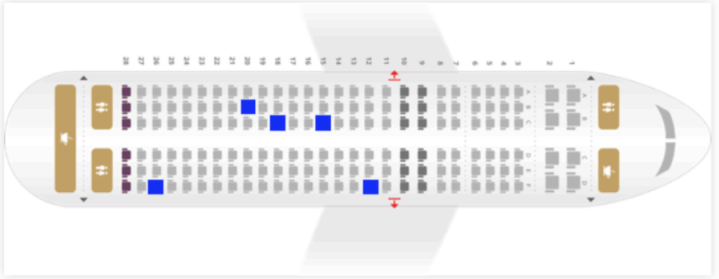
- ω = 7F, 16E, 21A
- $X(\omega) = 3$

Example 2: Airline no shows



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

Example 2: Airline no shows



- ω = 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.

Discrete random variables

A random variable is **discrete random variable** if its range, i.e. the set of values it can take, is a finite or countably infinite set.

Both our examples (dice, airline no-shows) were discrete random variables: you can count the outcomes on your fingers and toes. Something that can take on a continuous range of values (e.g. temperature, speed) is called a **continuous random variable**.

For now, we will focus exclusively on discrete random variables. (The math and notation for continuous random variables is more fiddly, but the concepts are the same).

Probability mass function

Suppose that X is a discrete random variable whose possible outcomes are some set $\mathcal{X}$. Then the probability mass function of X is

$$p_X(x) = P(X = x)$$

We always use X for the random variable, and x for some possible outcome.

Facts about PMFs:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in \mathcal{X}} p_X(x) = 1$
- For any set of numbers S, $P(X \in S) = \sum_{x \in S} p_X(x)$

Example: Probability mass functions

Dan, my friend from grad school, has just pulled up to his new house after a long cross country drive.

But he discovers that the movers have bailed and left all his furniture and boxes sitting in the front yard. What a bummer!

He decides to ask his new neighbors for some help getting his stuff indoors.

Assuming his neighbors are the kindly type, how many pairs of hands might come to his aid?

Example: Probability mass functions

Let's use a capital X to denote the (unknown) size of the household next door.

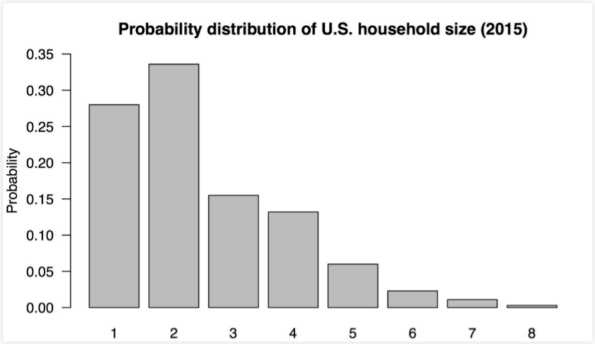
We'll let little x be some specific possible outcome.

Table: a probability mass function $p_X(x)$, taken from census data.

Size of household x	Probability, P(X = x)
1	0.280
2	0.336
3	0.155
4	0.132
5	0.060
6	0.023
7	0.011
8	0.003

Example: Probability mass functions

This is easier to visualize in a bar graph:



Example: Probability mass functions

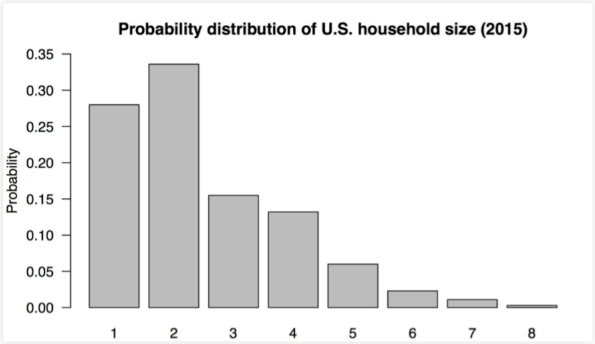
This probability mass function provides a complete representation of your uncertainty in this situation. It has all the key features of any probability distribution:

- An uncertain outcome: the composition of the family next door.
- A numerical function of that uncertain outcome: how large the family is.
- Finally, there are probabilities for each possible value of the random variable (here provided in a look-up table or bar graph).

Most probability distributions won't be this simple, but they all have these same three features.

Example: Probability mass functions

When you knock on the door, how many people do you “expect” to be living next door?



Expected value

The **expected value** or expectation of a random variable X is the **weighted average of the possible outcomes**, where the weight on each outcome is its probability.

Size x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Probability, P(X = x)	0.280	0.336	0.155	0.132	0.060	0.023	0.011	0.003

$$\mathbb{E}(X) = (0.280) \cdot 1 + (0.336) \cdot 2 + \cdots + (0.011) \cdot 7 + (0.003) \cdot 8 \approx 2.5\,.$$

The more likely numbers (e.g. 1 and 2) get higher weights than $1/8$, while the unlikely numbers (e.g. 7 and 8) get lower weights.

Expected value

This example conveys something important about expected values. Even if the world is black and white, an expected value is often grey:

- the “expected” American household size is 2.5 people
- a baseball player “expects” to get 0.25 hits per at bat
- you “expect” to get 0.5 heads per coin flip.
- you “expect” that your newborn child will have 0.97 ovaries (since 100/206 newborns are female).

Even if the underlying outcomes are all whole numbers, the expected value doesn’t have to be.

Expected value: Definition

Suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$ with probabilities $P(X = x_i)$.

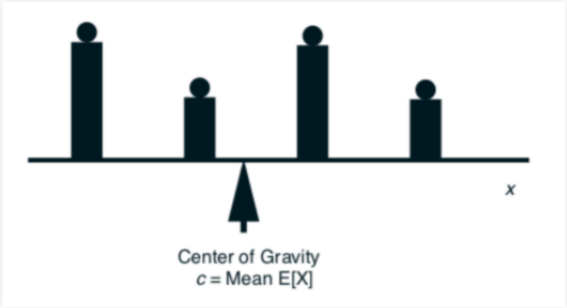
The formal definition for the expected value of X is

$$\mathbb{E}(X) = \sum_{i=1}^{N} P(X = x_i) \cdot x_i = \sum_{x \in \mathcal{X}} p_X(x) \cdot x$$

This measures the "center" or mean of the probability distribution.

Expected value: Definition

A nice interpretation here is that the expectation is like the center of gravity of a probability distribution. Arrange each outcome x_i along the real line, and place a mass equal to $P(X = x_i)$ at each point. The expected value is the center of gravity, i.e. where the masses would balance like on a see-saw:



Variance: Definition

A related concept is the **variance**, which measures the dispersion or spread of a probability distribution.

It is the expected (squared) deviation of a random variable from its mean:

$$\text{var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big)$$

The standard deviation of a probability distribution is $\sigma = \text{sd}(X) = \sqrt{\text{var}(X)}$.

The standard deviation is more interpretable than the variance, because it has the same units (dollars, miles, etc.) as the random variable itself.

Calculating the variance “by hand”

What’s the variance of the household size distribution?
(on your own time)

A useful identity for calculating variance is:

$$\text{var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Prove this!
(also on your own time)

Summarizing what we've done (and where we're going)

Started with **probability**
→ properties, conditioning, rules

From there we moved to **random variables**
→ functions of uncertain events, distributions, moments

Now, we'll discuss **probability models**
→ describing the real(uncertain)-world

Building probability models

A **probability model** is a stylized description of a real-world system in terms of random variables. To build one:

1. Identify the uncertain outcome of interest (e.g. a soccer game between UT and Texas Tech) and corresponding sample space Ω (e.g. all possible game scores).

2. Identify the random variables associated with that uncertain outcome, e.g. X_{UT} = the number of goals scored by UT, and X_{Tech} = the number of goals scored by Texas Tech. (Remember: a random variable is just a numerical summary of an uncertain outcome.)

3. Specify a probability distribution that lets us calculate probabilities associated with each random variable.

Building probability models

Step **3** – specify a probability distribution for the random variables of interest – is the hardest!

Imagine trying to list out all possible scores of UT-Tech and a probability for each one ...

Instead of building distributions from scratch, we will rely on a simplification called a **parametric probability model**.

Parametric models

A **parametric probability** distribution, or parametric probability model, is a probability distribution that can be completely described using a relatively small set of numbers.

These numbers are called the parameters of the distribution.

Lots of parametric models have been invented for specific purposes:

- normal, binomial, Poisson, *t*, chi-squared, Weibull, ...
- A large part of getting better at probability modeling is to learn about these existing parametric models: what they are, and when they're appropriate.

Parametric models

Suppose X is a discrete random variable with possible outcomes $x \in \mathcal{X}$. In a parametric model, the PMF of X takes the form

$$P(X = x) = f(x; \theta),$$

where θ is the parameter (or parameters) of the model.

The parameter θ completely specifies the PMF. Contrast this with the family-size example, where we had to write out the PMF explicitly for each possible outcome.

Pretty abstract!

Poisson distribution

The Poisson distribution models the number of "events" that occur in a pre-specified interval (e.g. time or space):

- How many goals will UT score in their game against Tech? (event = goal, interval = 90-minute game)
- How many couples will arrive for dinner at a hip new restaurant between 7 and 8 PM on a Friday night? (event = arrival of a couple, interval = one hour)
- How many irate customers will call AT&T customer service in the next minute? (event = phone call, interval = one minute)
- How many potholes will you encounter on a randomly chosen one-mile strip of I-35? (event = pothole, interval = one mile)

Poisson distribution

The possible outcomes of a Poisson random variable are the non-negative integers, $k = 0, 1, 2, \ldots$

The Poisson distribution is parametrized by a rate, conventionally denoted λ. (E.g. 2.1 goals per game, 3.4 potholes per mile, etc.)

The PMF of Poisson distribution takes the following form:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

The expected value of a Poisson random variable is the rate λ. The variance of a Poisson random variable is also the rate λ. (That's not a typo.)

Bernoulli distribution

Also called a “Bernoulli trial,” the possible outcomes are simple, either $B = 0$ or 1.

The Bernoulli distribution has only 1 parameter, $p \in [0, 1]$

The PMF of a Bernoulli random variable is pretty basic:

$$p_B(b) = \begin{cases} p & \text{if b=1} \\ 1 - p & \text{if b=0} \\ 0 & \text{otherwise} \end{cases}$$

Since B is Bernoulli, concise notation is $B \sim \text{Bern}(p)$

Binomial distribution

Combining Bernoulli's together gets you the binomial distribution

Examples:

- A coin is flipped ten times and it comes up heads (1) or tails (0) each time.
- A mortgage lender like UFCU has 100,000 borrowers, each of whom defaults (1) or not (0) on their payments.
- CNN conducts a poll of 1200 voters, asking each one whether they intend to vote for the incumbent (B=1) or not (B=0).

We are interested in the total number of "yes" (1) outcomes.

Binomial distribution

Assumptions:

$\rightarrow$ We have a sequence of N Bernoulli trials.

$\rightarrow$ For each of the N Bernoulli trials, the result is "yes" (1) with probability p and "no" (0) with probability $1 - p$. Each Bernoulli trial has the same probability of a 1.

$\rightarrow$ Each Bernoulli trial is independent of the others.

Let X be the number of 1's in the N-trial sequence. We say that X has a binomial distribution with parameters N and p: $X \sim \text{Binomial}(N, p)$.

Binomial distribution

In general, the PMF of a **binomial distribution** looks like this:

$$p_X(k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \binom{N}{k} = \frac{N!}{k!(N-k)!}$$

$\binom{N}{k}$ is the binomial coefficient, read aloud as “N choose k.”

Why?

Example: Airline no shows

Let's use the binomial distribution as a probability model for our earlier example on airline no-shows.

Say the airline sold tickets to 140 people, each of which will either show up to fly that day (1) or not (0).

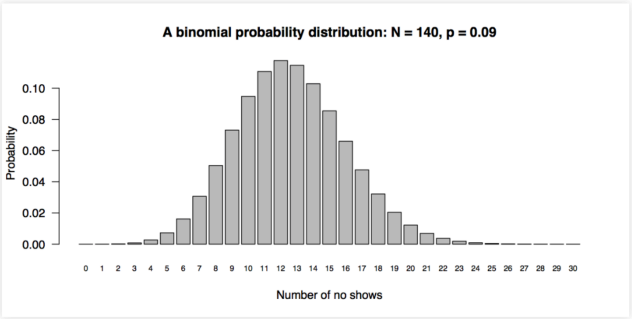
Our random variable is $X =$ number of no shows

Let's make two simplifying assumptions:
– Each person shows up independently of the other people.
– The probability of any one person failing to show up is 9%, the industry average.

Example: Airline no shows

Under these assumptions, $X \sim$ Binomial(N=140, p=0.09), with PMF

$$p_X(k) = \binom{140}{k} (0.09)^k \, (1 - 0.09)^{140-k}$$



Example: Airline no shows

Your turn! Suppose that:

The airline sold 140 tickets for 135 seats.

So if there are five no-shows, everyone gets a seat and the flight is full. You don't have a ticket, but you are 7th on the standby list. What is the probability of getting a seat under this model?
How would we calculate this probability?

Example: Airline no shows

To get a seat, you need 12 people to no-show (5 oversolds seats + 7 standby passengers).

To calculate $P(X \geq 12)$, we add up all the probabilities for $X = 12$, $X = 13$, $X = 14$, all the way up to $X = 140$.

This is called an "upper tail area" of the PMF.

$$P(X \geq 12) = \sum_{k=12}^{140} p_X(k) = \sum_{k=12}^{140} \binom{140}{k} (0.09)^k \, (1 - 0.09)^{140-k}$$

(see `R` script `probability.R`)

Distribution and density functions

Up to this point, we've talk about probability mass functions (PMFs), also known as distribution functions

We also have *integrals* of distribution functions – these are called density functions or cumulative distribution functions (CDFs)

Example (from previous slide):

$P(X \geq 12) = 1 - P(X \leq 11) = \sum_{k=12}^{140} p_X(k) =$ "1 minus the cumulative distribution function of a Poisson variable evaluated at 11."

Making this more formal

The cumulative distribution function, or CDF, is defined as:

$$F_X(x) = P(X \leq x)$$

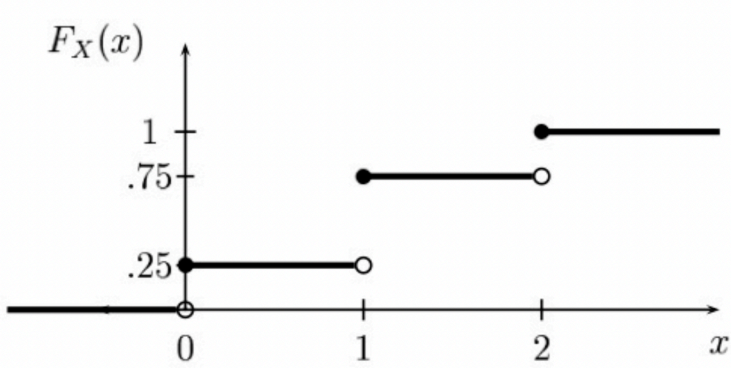
Facts:

All random variables have a CDF.

The CDF completely characterizes the random variable: if X and Y have the same CDF, then for all sets S, $P(X \in S) = P(Y \in S)$.

If this holds, we say that X and Y are equal in distribution. This doesn't mean they're identical! It just means that all probability statements about X and Y will be identical.

CDF: Simple example



The CDF for Binomial(N=2, p=0.5). (Let's write this on the board.) The jumps correspond to the points where the PMF has positive probability. What is $F(1)$? What is $F(0.6)$? What is $F(17)$?

PDF-CDF relationship

Note that, by the definition of the CDF and PDF, we have the following relationship:

$$F_X(x) = P(X \in (-\infty, x)) = \int_{-\infty}^{x} f_X(x)dx$$

Remember the Fundamental Theorem of Calculus! This relationship says that the PDF is the derivative of the CDF:

$$f_X(x) = F_X^{'}(x)$$

at all points where $F_X(x)$ is differentiable.

Summary table for R implementation

	distribution	PMF/PDF	CDF	simulate
Discrete	Bernoulli	`dbinom()`	`pbinom()`	`rbinom()`
	Binomial	`dbinom()`	`pbinom()`	`rbinom()`
	Poisson	`dpois()`	`ppois()`	`rpois()`
Continuous	Normal	`dnorm()`	`pnorm()`	`rnorm()`
	Uniform	`dunif()`	`punif()`	`runif()`
	Exponential	`dexp()`	`pexp()`	`rexp()`

Commonly used distributions and how to implement in **R**!