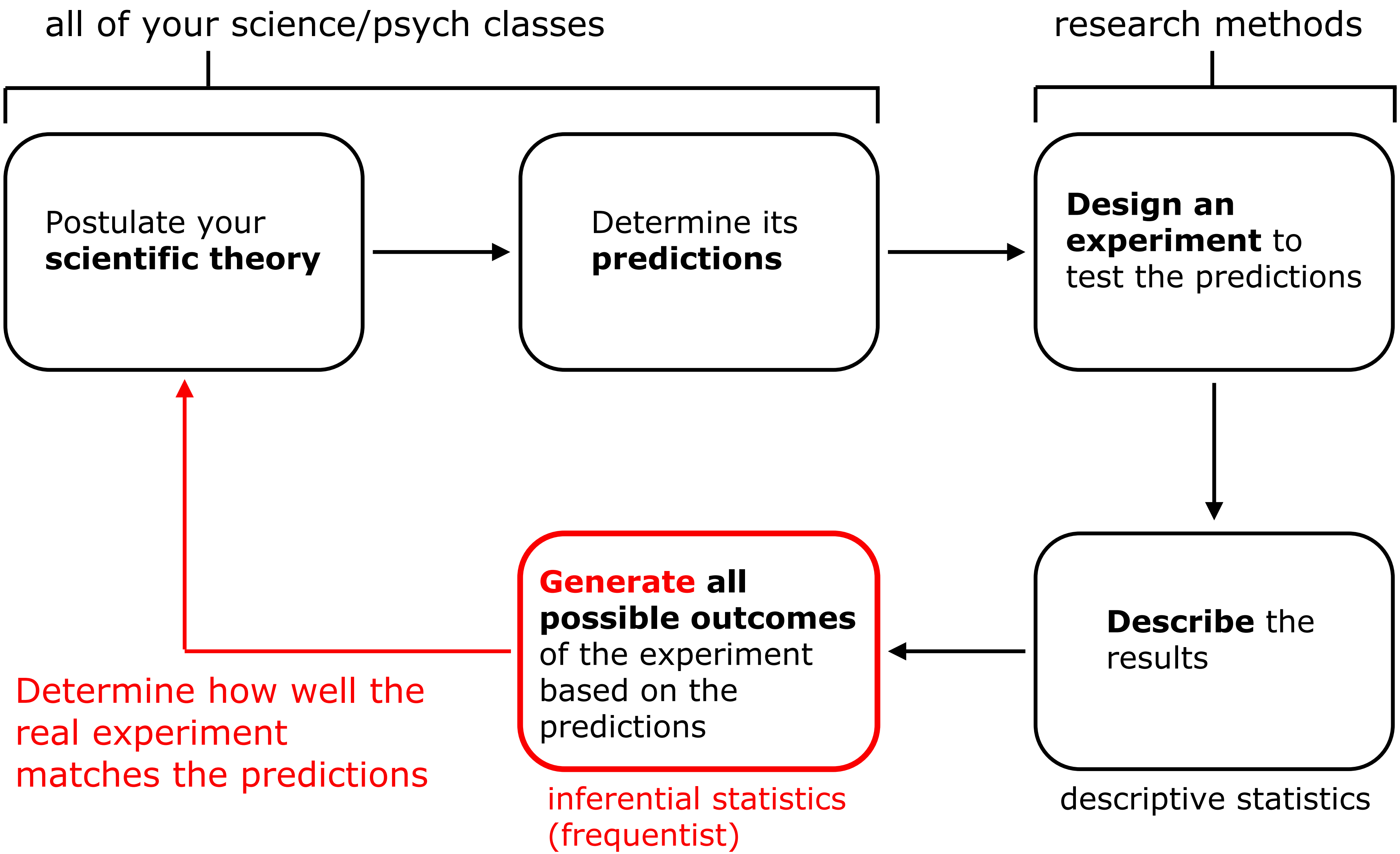
جامعة نيويورك أبوظبي
NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

Class 7: Foundations of null hypothesis testing - probability and falsification

Prof. Jon Sprouse
Psychology

The next step: inferential statistics



Probability Basics

(frequentist)
(You are already good at this!)

Probability

Probability is a mathematical statement about how likely an event is to occur. It takes a value between 0 and 1, where 0 means the event will never occur, and 1 means the event is certain to occur. (You can also think of it as a percentage 0% to 100%, but to be precise, probability ranges from 0 to 1.)

Probability as long-run relative frequency:

As we have already seen, the frequentist (or objective) approach to probability defines it as the relative frequency of the event after a large number of repetitions (ideally approaching an infinite number of repetitions):

$$P(\text{event}) = \frac{\text{frequency of the event}}{\text{\# of repetitions}}$$

This means "probability of the event"

If the possible outcomes are exhaustive and equally likely, you can calculate probability without a simulation. It is just the proportion of critical outcomes to all possible outcomes:

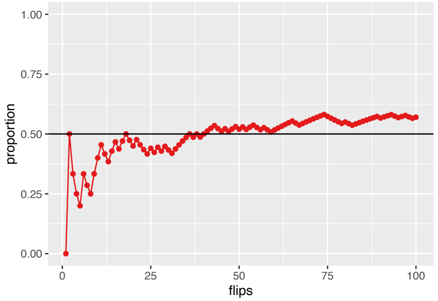
$$P(\text{event}) = \frac{\text{critical outcomes}}{\text{all possible outcomes}}$$

Probability - start with a coin

Flipping a coin:

If the coin is fair, then the two outcomes are equally likely. This means we can either simulate the long-run frequency or calculate the probability directly:

$$P(\text{heads}) = \frac{1}{2}$$



Heads or Tails
(one flip)

If two events are mutually exclusive (that means they cannot both occur simultaneously), then we can calculate the probability of either occurring with addition:

$$P(\text{heads or tails}) = P(\text{heads}) + P(\text{tails}) = \frac{1}{2} + \frac{1}{2} = 1$$

Heads and Tails
(two flips or two coins!)

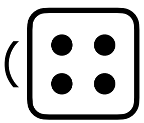
If two events are independent (that means they don't influence each other), then we can calculate the probability of both occurring with multiplication:

$$P(\text{heads and tails}) = P(\text{heads}) \text{ x } P(\text{tails}) = \frac{1}{2} \text{ x } \frac{1}{2} = \frac{1}{4}$$

Probability - then try a six-sided die

6-sided die:

If the die is fair, then the two outcomes are equally likely. This means we can calculate the probability directly:

P() $= \frac{1}{6}$

2 or 4
(one roll)

If two events are mutually exclusive (that means they cannot both occur simultaneously), then we can calculate the probability of either occurring with addition:

$$P(2 \text{ or } 4) = P(2) + P(4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

2 and 4
(two rolls or two dice!)

If two events are independent (that means they don't influence each other), then we can calculate the probability of both occurring with multiplication:

$$P(2 \text{ and } 4) = P(2) \times P(4) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Probability - then try a deck of cards

There are 52 possible cards. There are 4 “suits” - hearts, diamonds, clubs, spades. There are 13 values - 2 through 10 and the Jack, Queen, King, and Ace.

	2	3	4	5	6	7	8	9	10	J	Q	K	A
♠	•	•	•	•	•	•	•	•	•	•	•	•	•
♣	•	•	•	•	•	•	•	•	•	•	•	•	•
♦	•	•	•	•	•	•	•	•	•	•	•	•	•
♥	•	•	•	•	•	•	•	•	•	•	•	•	•

What is the probability of drawing a jack?

$$P(J) = \frac{\text{number of events you care about}}{\text{total number of events}} = \frac{4}{52} \cong .08$$

And what is the probability of drawing a heart?

$$P(\heartsuit) = \frac{\text{number of events you care about}}{\text{total number of events}} = \frac{13}{52} = .25$$

Probability - then try a deck of cards

We see the same properties for OR and AND:

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**J**	**Q**	**K**	**A**
♠	•	•	•	•	•	•	•	•	•	•	•	•	•
♣	•	•	•	•	•	•	•	•	•	•	•	•	•
♦	•	•	•	•	•	•	•	•	•	•	•	•	•
♥	•	•	•	•	•	•	•	•	•	•	•	•	•

What is the probability of drawing a jack or a 7?

$$P(J \text{ or } 7) = P(J) + P(7) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13} \cong .15$$

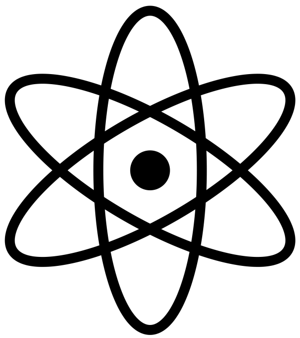
What is the probability of drawing a jack and a 7?

$$P(J \text{ and } 7) = P(J) \text{ x } P(7) = \frac{4}{52} \text{ x } \frac{4}{52} = \frac{1}{169} \cong .006$$

Conditional Probability

(This will feel new)

Why do we care?



In the inferential statistical methods that we are about to learn, called Null Hypothesis Testing, we calculate a specific **conditional probability**:

P(data | null hypothesis)

This conditional probability tells us the probability of the data given the assumption that the null hypothesis is true.

This is the metaphorical heart of Null Hypothesis Testing. So we need to understand how conditional probabilities work, and what they tell us.

The critical concept: conditional probability

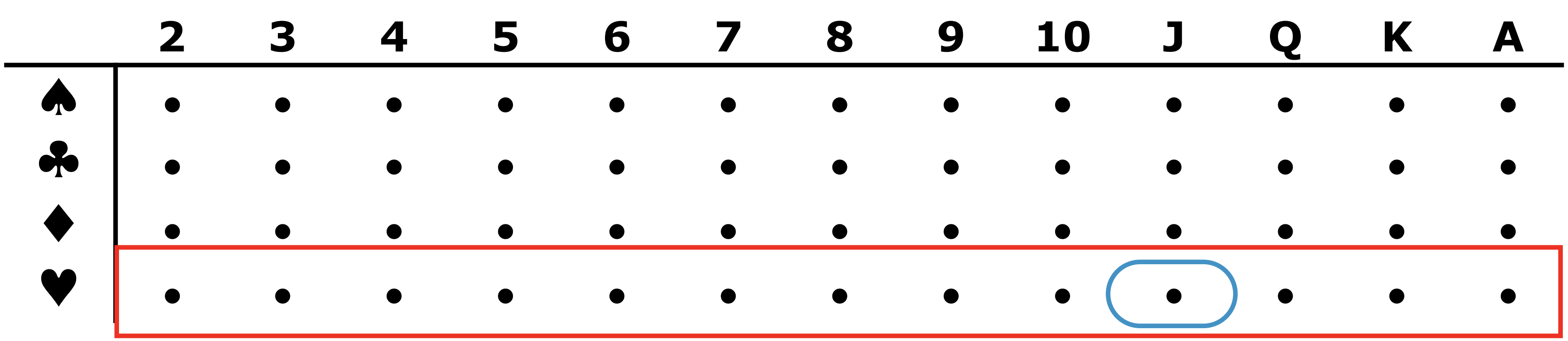
Conditional Probability: The probability of an event **given that** another event has occurred.

Let's say you draw a card, but can't see it. Your friend tells you it is a heart. What is the probability that it is a Jack?

This is a conditional probability. It is asking what the probability of a Jack is given that the card is a heart.

$$P(J \mid \heartsuit) = \frac{\text{number of events that are both Jack and heart}}{\text{number of heart events}} = \frac{1}{13}$$

The pipe symbol means "given that"



The way to think about it - focus!

Conditional probability is about **focusing attention on the world created by the given event**.

If you look at the definition of conditional probability, the "given" event is in the denominator. The denominator (division) is the mathematical way of saying "focus only on these events".

$$P(J \mid ♥) = \frac{\text{number of events that are both Jack and heart}}{\text{number of heart events}} = \frac{1}{13}$$

focus on hearts

We also see this in the definition of basic probability. The denominator is the set of all of the events (both coin faces, all six die faces, the 52 cards). This is telling us to focus on all outcomes. That is what basic probability is!

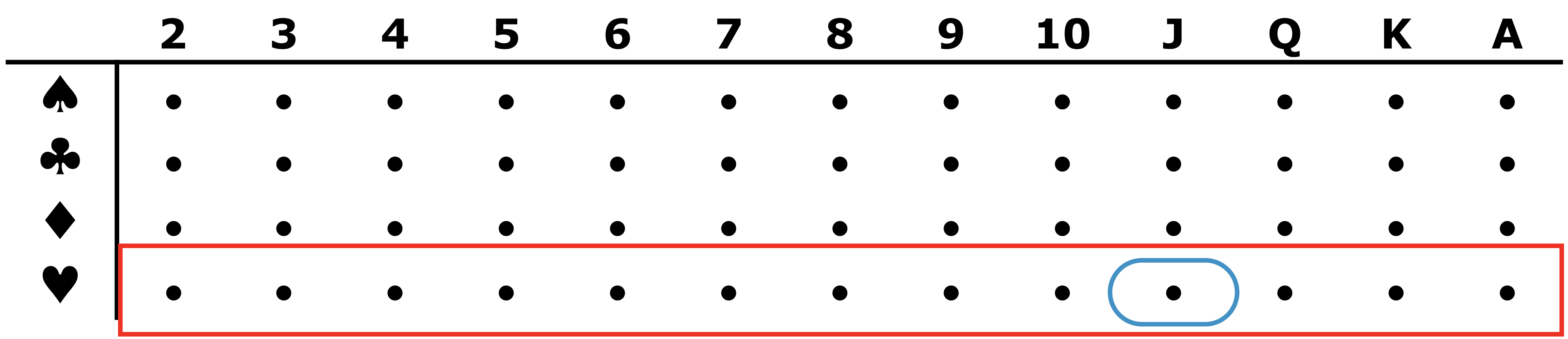
$$P(\text{event}) = \frac{\text{critical outcomes}}{\text{all possible outcomes}}$$

focus on all outcomes

The way to think about it - focus!

Conditional Probability: The probability of an event **given that** another event has occurred.

$$P(J \mid \heartsuit) = \frac{\text{number of events that are both Jack and heart}}{\text{number of heart events}} = \frac{1}{13}$$



The denominator tells us to focus only on the heart events. That is our space of possibilities.

The numerator tells us to look for the Jack event within the space of possibilities (Jack and heart).

The order of conditional probabilities

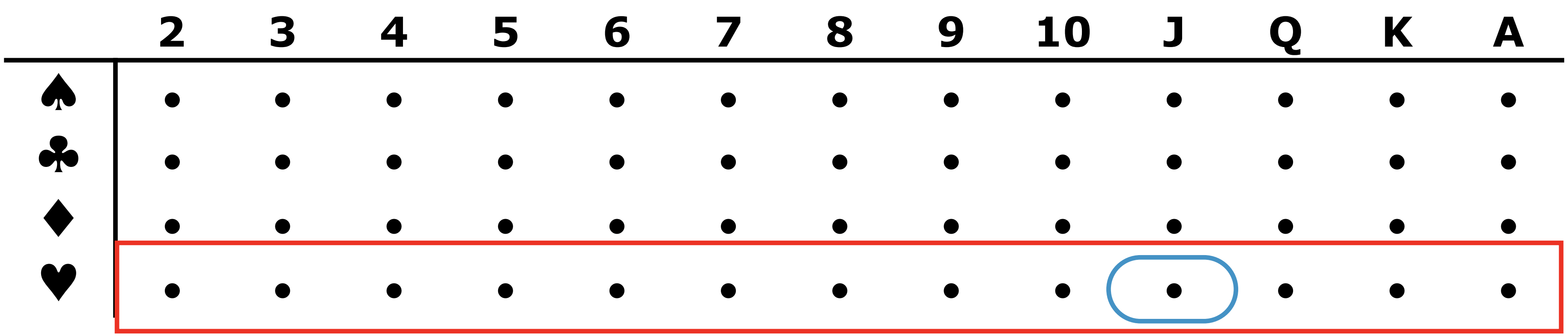
(This is important!)

The order of the conditional really matters!

The “given” (after the pipe) tells us the space of possible outcomes. This really changes the meaning of conditional probabilities

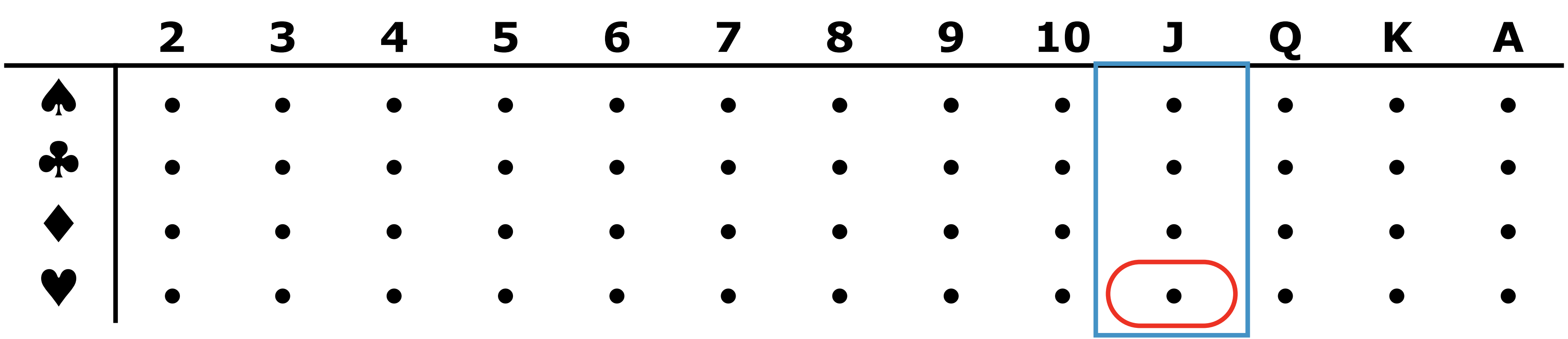
What is the probability of a Jack given that the card is a heart?

$$P(J \mid \heartsuit) = \frac{1}{13}$$



What is the probability of a heart given that the card is a Jack?

$$P(\heartsuit \mid J) = \frac{1}{4}$$



An example you might have intuitions about

Here is a fact - a lot of movie stars in the US live in Los Angeles. This is because much of the US movie industry is based there.

What is the probability of being a movie star given that someone lives in Los Angeles?

P(movie star | live in LA)

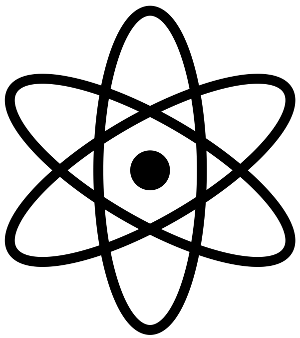
$$\frac{\text{number of movie stars in living in LA}}{\text{number of people living in LA}} = \frac{400}{4{,}000{,}000} = \mathbf{\sim .01\%}$$

What is the probability of living in LA given that someone is a movie star?

P(live in LA | movie star)

$$\frac{\text{number of movie stars in living in LA}}{\text{number of movie stars}} = \frac{400}{500} = \mathbf{\sim 80\%}$$

Why do we care?



In the inferential statistical methods that we are about to learn, called Null Hypothesis Testing, we calculate a conditional probability:

P(data | null hypothesis)

This conditional probability tells us the probability of the data given the assumption that the null hypothesis is true.

This is the metaphorical heart of Null Hypothesis Testing. So we need to understand how conditional probabilities work, and what they tell us.

Probability distributions

(From frequency to probability)

Probability Mass

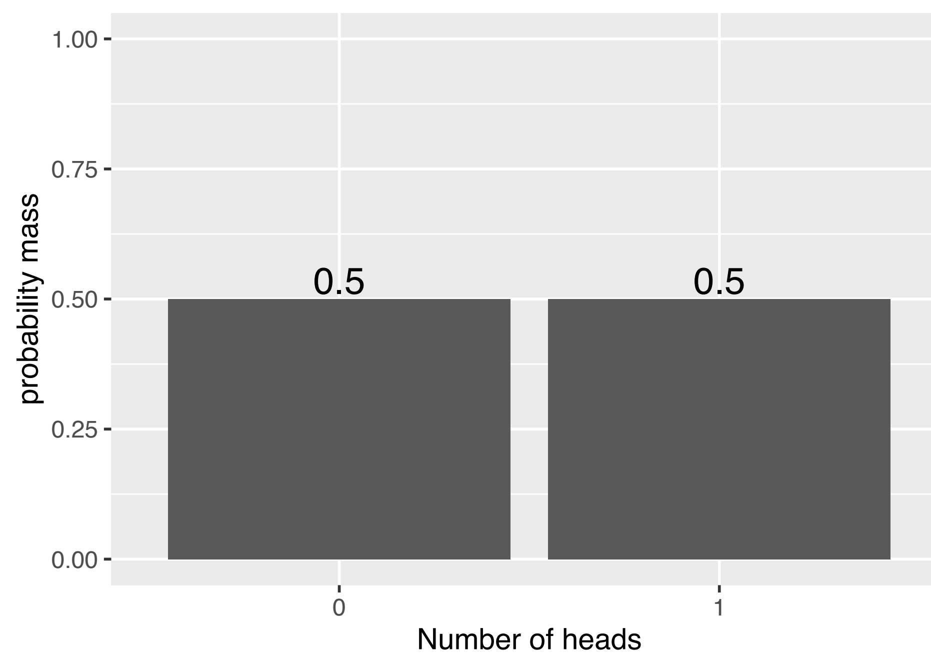
(Discrete Variables)

From frequency distributions to probability distributions

We have been working extensively with frequency distributions. We know them inside and out. Now here's the big reveal — we can also use distributions to explore probability.

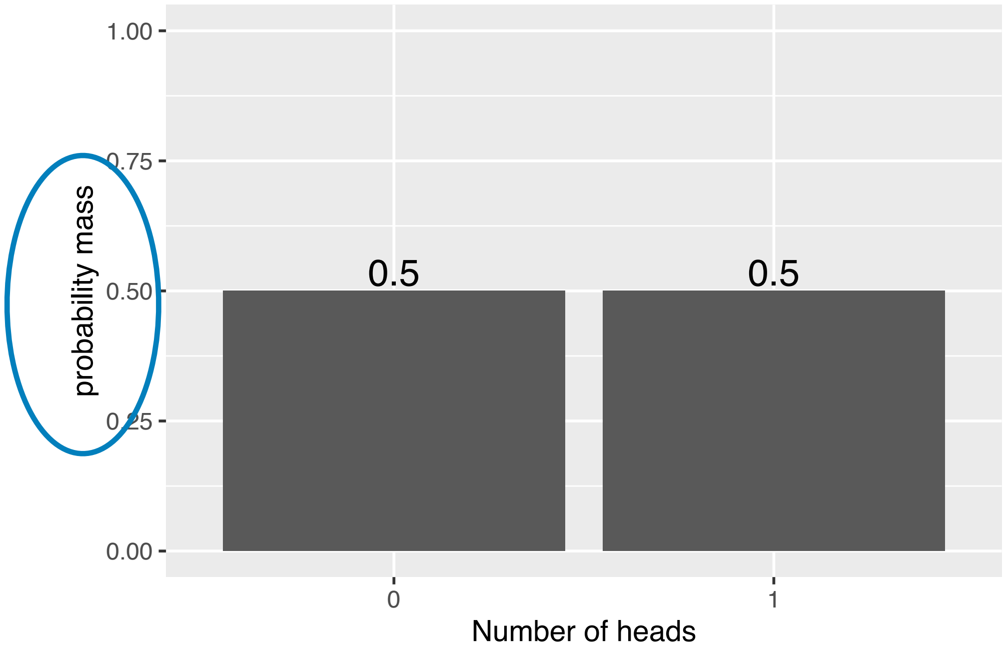
(We've already done this with the normal distribution. So this shouldn't be surprising. Also, probability is frequency under the frequentist view, so it shouldn't be surprising that we can use the same math for both frequency and probability!)

Here is a probability distribution for flipping 1 coin, and asking the probability that we get 0 heads (which is tails!) or 1 heads. It looks very similar to a frequency distribution.



The Probability Mass metaphor

The only difference in the plot between this probability distribution and a frequency distribution is the y-axis. For this probability distribution, which captures discrete events, the label on the axis is **probability mass**.



Probability mass is a metaphor. Mass just means "stuff". (Mass is one of those deep, unexplained things in physics. We really don't know what it is. It is the existence of matter.)

The idea behind this metaphor is that we can think of probability as stuff that can be distributed across the outcomes.

If you list all possible outcomes (within a given world that we focus on), the probability mass will sum to 1. (Like saying 100% of the probability mass!) This is an axiom of probability. It is part of the definition of probability. The probabilities of all of the (mutually exclusive) events will always sum to 1.

Probability mass with coin flips

To see probability mass at work, let's expand our example. Let's flip 1, 2, and 3 coins. First, we can look at the individual events. The probabilities for each outcome will add up to 1:

1 coin

= .5

= .5

2 coins

= .25

= .25

= .25

= .25

3 coins

= .125

= .125

= .125

= .125

= .125

= .125

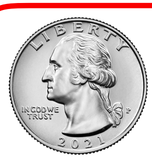
= .125

= .125

Probability mass with coin flips

To see probability mass at work, let's expand our example. Let's flip 1, 2, and 3 coins. First, we can look at the individual events. The probabilities for each outcome will add up to 1:

1 coin

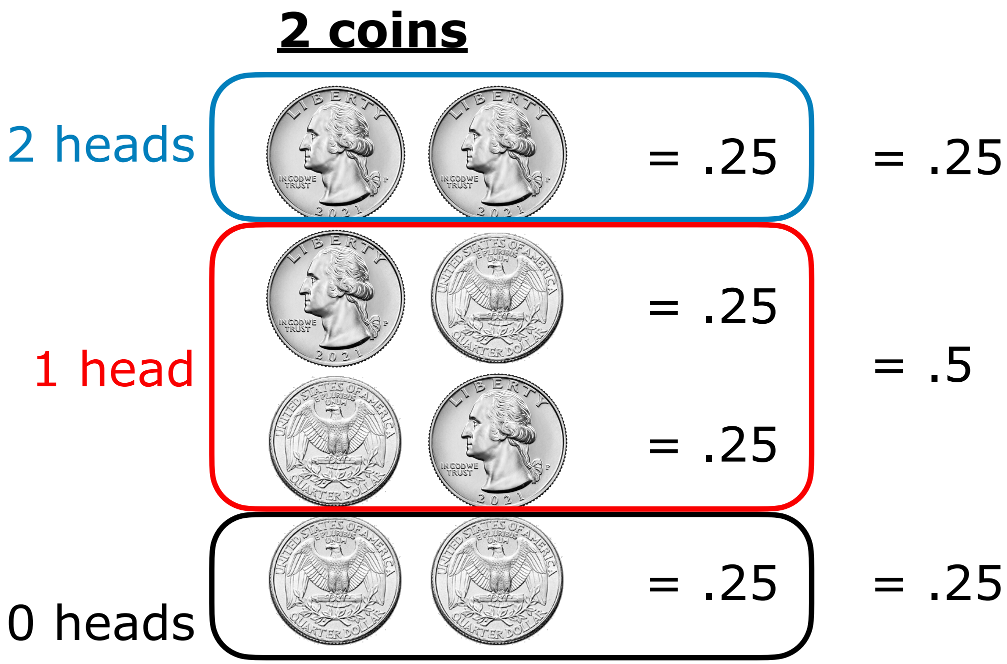


= .5 1 head

= .5 0 heads

Probability mass with coin flips

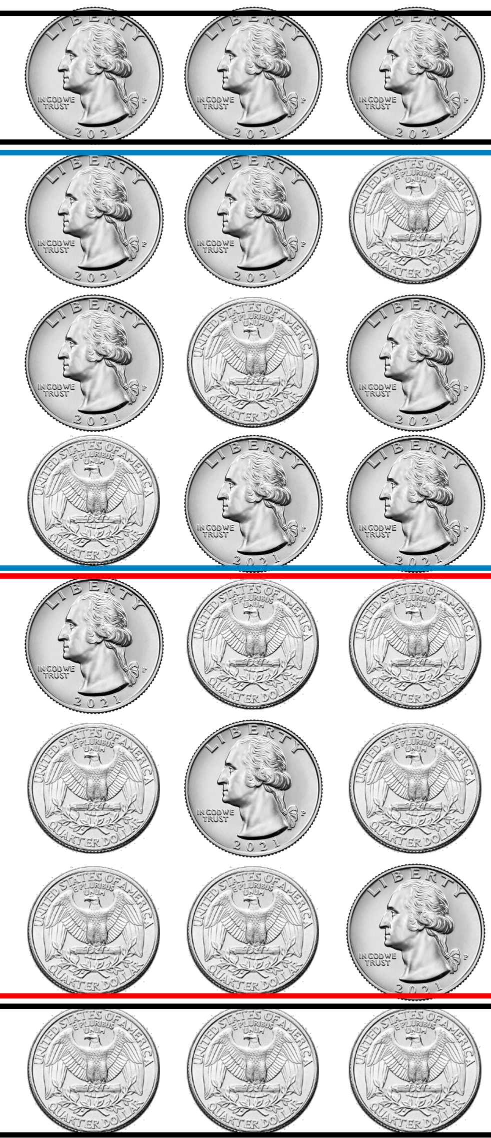
To see probability mass at work, let's expand our example. Let's flip 1, 2, and 3 coins. First, we can look at the individual events. The probabilities for each outcome will add up to 1:



Probability mass with coin flips

To see probability mass at work, let's expand our example. Let's flip 1, 2, and 3 coins. First, we can look at the individual events. The probabilities for each outcome will add up to 1:

3 coins

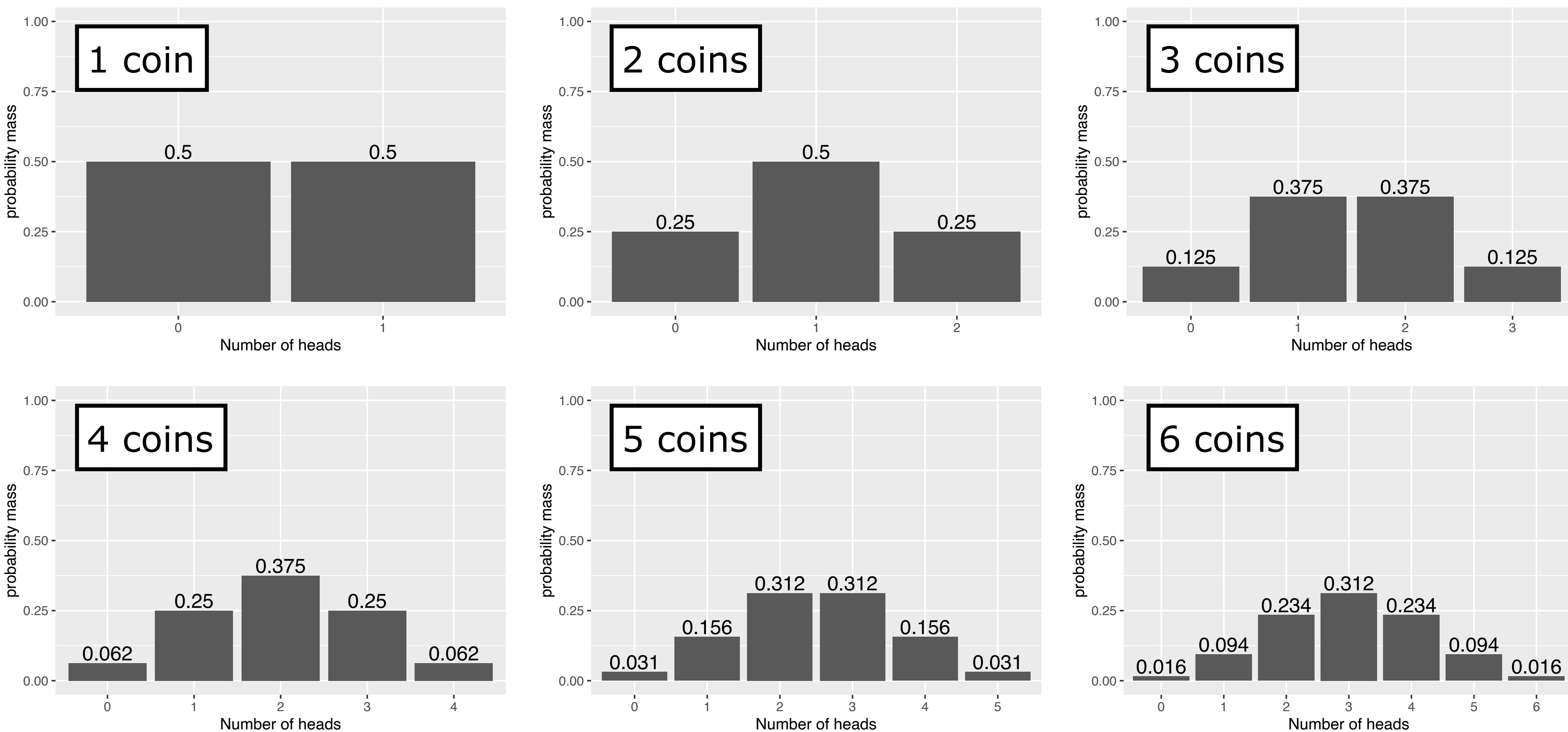


= .125	3 heads	= .125
		= .125
= .375	2 heads	= .125
		= .125
		= .125
= .375	1 head	= .125
		= .125
= .125	0 heads	= .125

Probability mass with coin flips

We can also look at these as distributions. It would be boring to simply plot the equal probabilities of the individual events (a bunch of bars of exactly the same height).

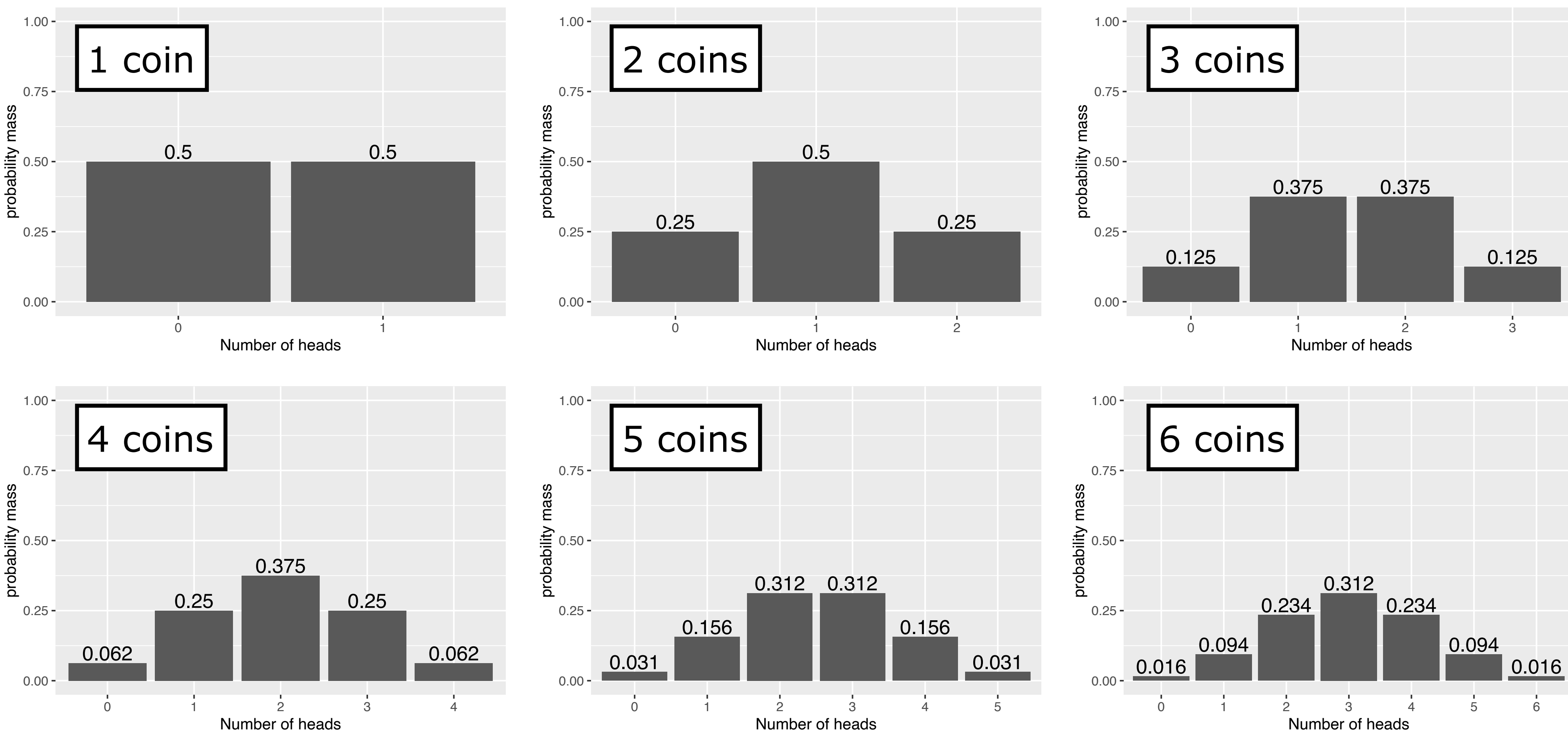
So, instead, let's plot the probability of getting 0-6 heads when we do 1-6 coin flips.



Probability mass with coin flips

First, notice that the probabilities of the events always sum to 1, even when we count number of heads.

Second, you will probably notice a pattern to the shapes. That is because coin flips follow a family of distributions called the **binomial distribution**. (We won't study it in this class, but it is worth knowing that it exists!)



Probability Density

(Continuous Variables)

Continuous variables

So far, we have only looked at <u>discrete</u> variables. Things get more complicated when we look at continuous variables.

We cannot simply state the probability (mass) for a given score on a continuous variable. This is because continuous variables, by definition, have an infinite number of scores.

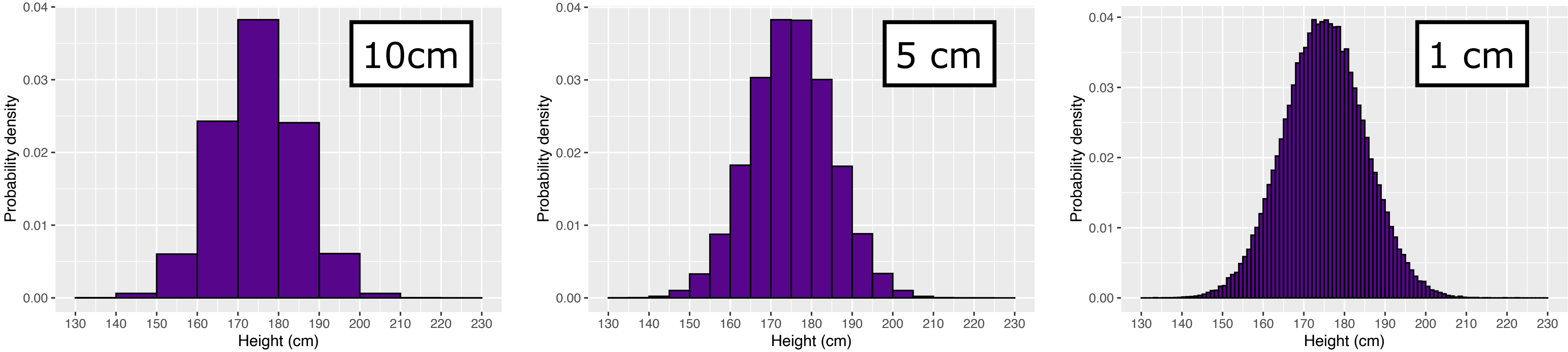
This means that the probability associated with any individual score is zero **because any number divided by infinity is zero**. (This is a mathematical fact.) If we plug infinity into the probability calculation, we will always get zero.

$$P(\text{event}) = \frac{\text{critical outcomes}}{\text{all possible outcomes}} = \frac{\text{critical outcomes}}{\infty} = 0$$

Infinity is weird. Our brains are not able to truly comprehend it. One way to get part way there is to imagine two things: (i) that the number of possible scores is really, really big; and (ii) that any specific score will only occur once — because we can always increase the precision of the measurement to make sure it is distinct from others (e.g., one height measurement of 172.5 might actually be 172.5837<u>5</u>, and a second measurement might be 172.5837<u>6</u>). This gets us very very small probabilities… which is part way to zero.

Continuous variables and bins

The solution is to create **bins** — just like we have already been doing in our distribution plots. This allows us to add up the probability mass in each bin so that it is not zero.



But now we are doing something a little different. The y-axis is no longer probability mass, which is probability for a single value. It is **probability for a range of values**. We call this probability density:

probability density: $\frac{\text{mass}}{\text{range}}$ The amount of probability in a specific range (bin) on a continuous scale.

The probability density metaphor

mass:		Stuff.
density:	$\frac{\text{mass}}{\text{volume}}$	The amount of stuff in a specific volume. (Things with more stuff in the same amount of volume have higher density. Lead has higher density than water.)

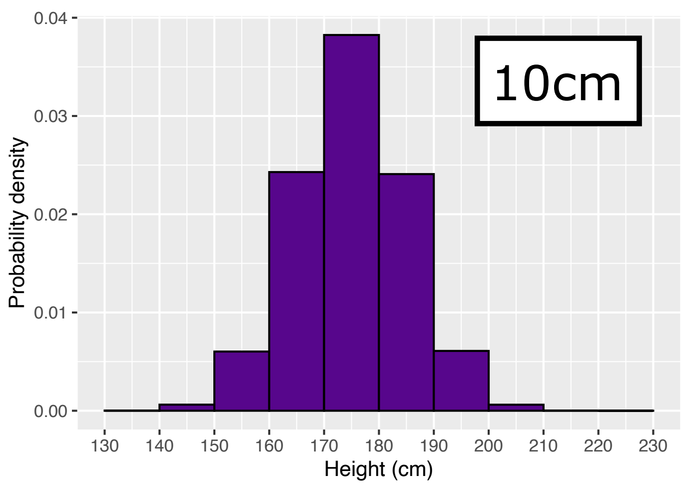
probability mass:		Probability as (metaphorical) stuff.
probability density:	$\frac{\text{mass}}{\text{range}}$	The amount of probability in a specific range (bin) on a continuous scale. (Events that are more probable in the same sized range on the scale will have higher probability density.)

Interpreting probability density

Once you see the metaphor, it is straightforward to interpret probability density. You can take the width of the bin and multiply it by the density to get the probability mass of obtaining a score in the range of the bin. This follows from algebra and the equation for density:

$$\text{density} = \frac{\text{mass}}{\text{volume}}$$

$$\text{mass} = \text{density} \times \text{volume}$$



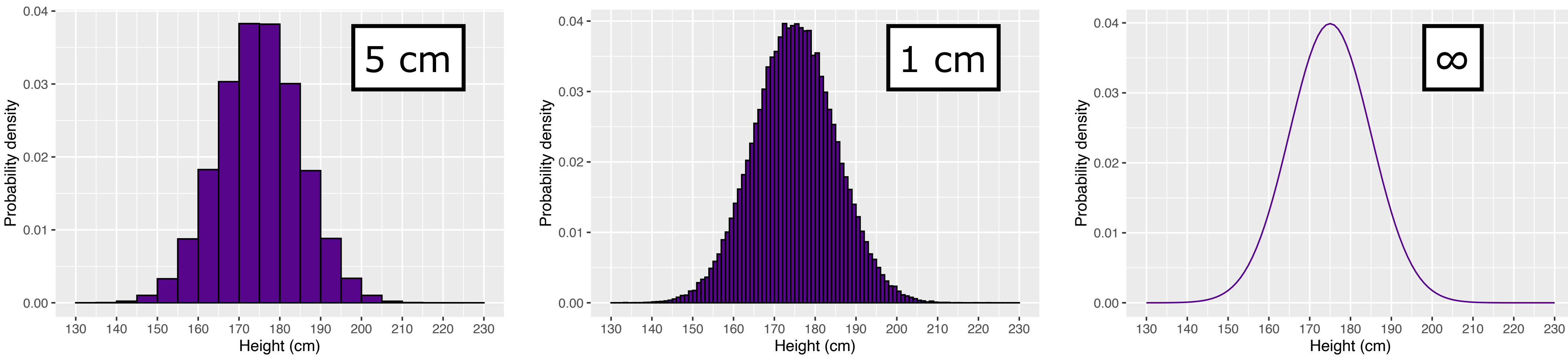
$$\text{probability density} = \frac{\text{probability mass}}{\text{range}}$$

$$\text{probability mass} = \text{probability density} \times \text{range}$$

The 160-170 bin has a density around .025. So the probability mass is around .025*10 = .25. The 170-180 bin has a density around .0375, so the probability mass is around .0375*10 = .375. And if you do this for each of the bars, then sum their probability masses, they will <u>add up to 1</u>.

Smaller and smaller bins

You may recall from calculus that you can make smaller and smaller bins, approaching infinitesimally small bins, until you end up with a smooth curve.



We can use the same logic with these smooth curves - the probability mass can be calculated from the probability density and a range of values.

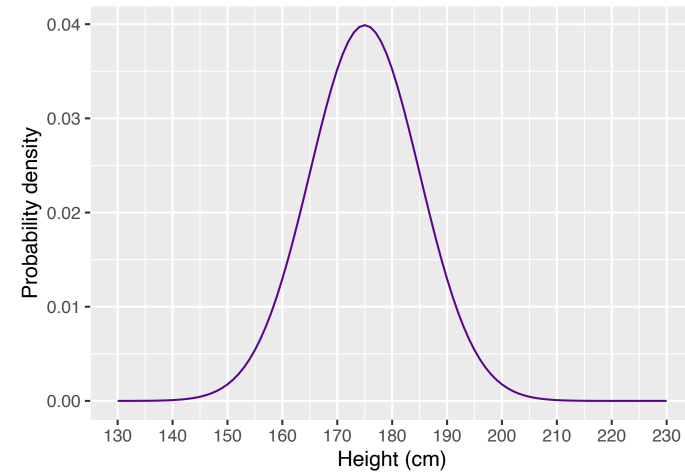
(This is also a case where knowing some calculus can be illuminating. We just ran into the idea of integrals/integration **in a real issue in our work!** We won't pursue it here because I am not an expert on calculus. But if you've ever wondered if calculus would show up in your life, it does, at least in bits and pieces, if you pursue science.)

Probability Density Functions and the area under the curve

(Our major tool in inferential statistics!)

Probability density functions

Now that we understand what probability density is. We can connect one more concept.



Remember that smooth distributions like this are generated by equations. Those equations are **functions** - they relate one quantity (the x value) to another (the y value).

The functions that define probability distributions are called **probability density functions** because the value that they yield, which I have been calling f(x) to be as generic as possible, is **probability density**.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

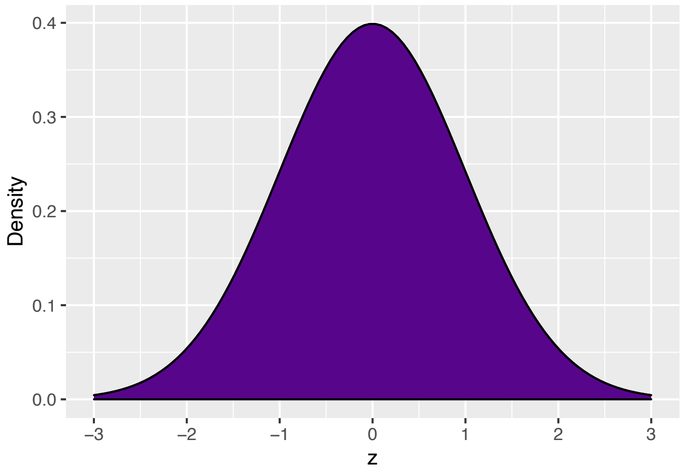
probability density

There is nothing too deep about this that you need to do or memorize. I just want you to know this term because you will see it whenever you read any books or papers about statistics. And you now have all of the knowledge you need to see why they are called probability density functions.

Area under the curve is probability

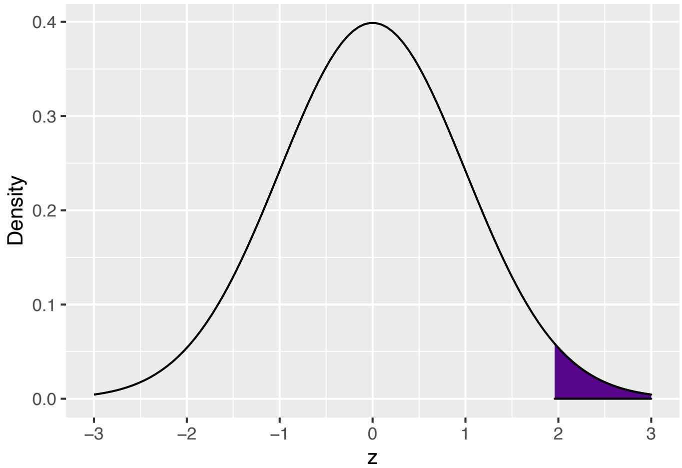
The area under the curve of a probability density function is 1.

This follows from the fact that probability mass sums to 1, and the fact that the probability density function shows the full range of the scale. (Basically, it is like having a single bin that takes up the full range!)



This leads to one more shift in the way we think. Instead of thinking about percentile rank (an ordinal measure of frequency), we can think about the probability of the range of scores.

The shaded region marks the area under the curve between a z-score of 1.96 and the right edge. That represents a probability mass of .025 (from Table A1 or the pnorm() function).



In other words, the probability of observing a z-score of 1.96 or higher is .025. That means there is a 2.5% chance of observing a score this high or higher.

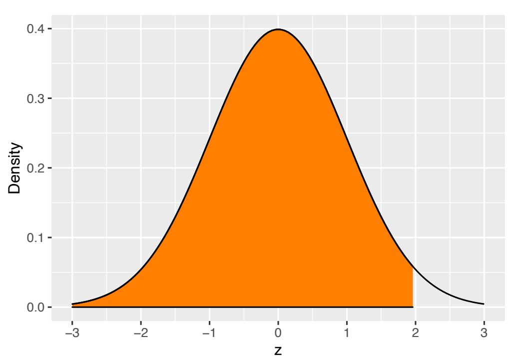
Percentile rank versus probability

Percentile rank and probability calculations use the <u>same math</u>.

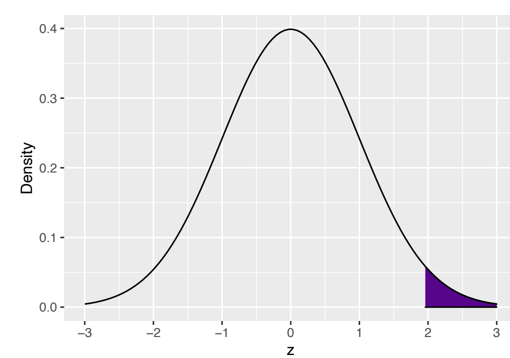
Percentile rank and probability calculations will yield <u>the same value</u> (though one is 0-100 and the other is 0-1).

The difference is in the <u>concept</u>.

Percentile rank is a frequency concept. It tells us how many scores are above or below a critical value.



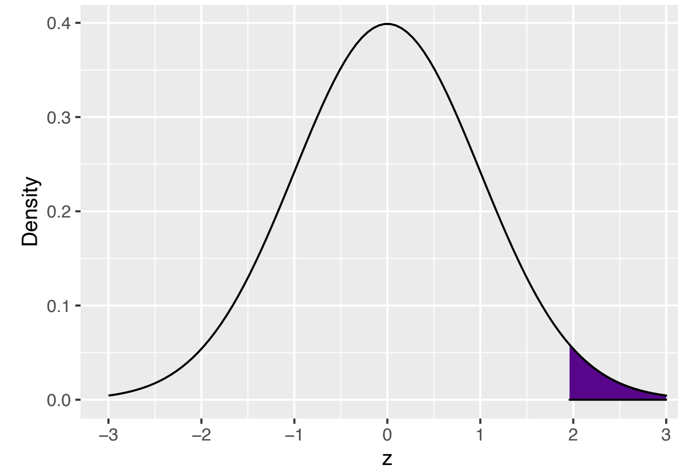
Probability is its own thing. It tells us how likely we are to get a score above or below a critical value



It is worth remembering that frequency has only one definition. It is the count of something. Probability has two different definitions: long-run relative frequency (frequentist) or subjective belief (Bayesian). This helps to show that they are different concepts (even if they are intimately related, particularly under the frequentist view).

This is the foundation of inferential statistics

Given a probability distribution and a score, we can now calculate the probability of obtaining that score or one more extreme.



We can do this either with R functions like pnorm() or with z-scores and the table in our book.

This is the basis of the inferential statistical tests that we are about to learn. These tests calculate the probability of obtaining the the data from our experiment or data more extreme.

This is a major tool for us as scientists! Remember, we want to know how extreme our experimental result is. Now we can answer that question — we can say the probability of obtaining our result or a result more extreme!

Falsification and the Philosophy of Science

The Problem of Confirmation

Humans tend to want to <u>prove</u> the hypotheses that they are testing. (I certainly do!) We call this confirmation in the philosophy of science:

confirmation: The act of collecting evidence that <u>supports</u> a hypothesis.

But there is a major logical flaw with confirmation. So this very human bias tends can be a problem for science.

The problem of confirmation: Any given piece of evidence that can be used to confirm one theory can be used to <u>confirm an infinite number of theories</u>.

To see this in action, let's take a really simple theory: **All ravens are black** 

And let's say you set out to confirm this theory. Well, evidence that supports a theory is sometimes called <u>positive</u> evidence. It is evidence that is directly predicted by the theory. In this case, that would be black ravens. So you go out to find ravens, and see if they are black:

The Problem of Confirmation

Let's say that you are able to collect **300 ravens**, and **all of them are black**.

That is positive evidence for the theory that **all ravens are black**.

The problem is that it is also positive evidence for an infinite number of other theories that also predict at least 300 black ravens:

300 ravens are black and all of the rest are white.

301 ravens are black and all of the rest are white.

302 ravens are black and all of the rest are white.

303 ravens are black and all of the rest are white.

304 ravens are black and all of the rest are white.

305 ravens are black and all of the rest are white.

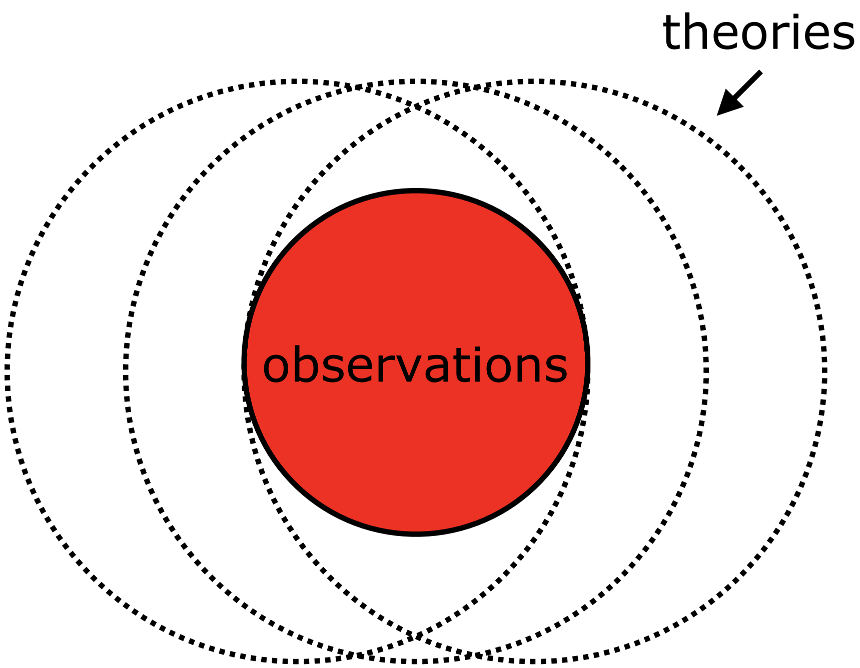
… and on and on…

I know this feels like a cheat, but these toy examples are there to show us the logic in its purest form. The same logic holds for complex theories too.

The Problem of Confirmation

The problem of confirmation: Any given piece of evidence that can be used to confirm one theory can be used to confirm an infinite number of theories.

One way to schematize this is by viewing the observations in our test as a subset of the predictions of the theory.



The problem of confirmation arises because different theories can make overlapping predictions. If the observation that we have is part of that overlap, we can't use it to uniquely confirm one theory.

If the evidence you have confirms an infinite number of theories, then you really haven't made progress at all.

One solution to the problem of confirmation: Falsification

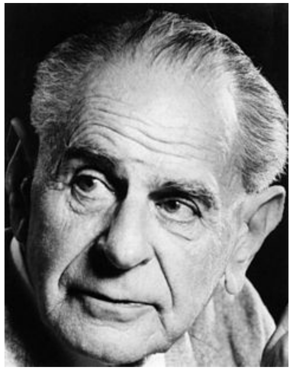
Falsification

Falsification: The act of collecting evidence that disproves a hypothesis.

The evidence that disproves a theory is sometimes called negative evidence. What would be negative evidence for the theory that **all ravens are black**?

A white raven! (Or any non-black color.)

Negative evidence is what the theory **does not predict**. In other words, in order for a theory to be testable, it must be clear what it doesn't predict. (This should remind you of our second class - "a theory that predicts everything predicts nothing"!)



Karl Popper
1902-1994

Falsification is most famously associated with Karl Popper, a philosopher who argued that "confirmation is a myth."

If a prediction is shown to be false, then the theory is falsified.

If a prediction is shown to be true, then we can't say anything about the theory. All we can say is that the theory has not yet been falsified.

Falsification is considered the "gold standard" by many scientists

The actual process of science is messy. It is rare to have a theory with such clear predictions as "all ravens are black", and it is rare to have tests that are as clear cut as "find a white raven".

Nonetheless, many scientists subscribe to falsification. They strive to build theories that make clear predictions, and they strive to create tests that look for negative evidence (something the theory does not predict).

Why do we care?

The inferential method in frequentist statistics is called **Null Hypothesis Significance Testing**. It is predicated upon the idea of falsifying a hypothesis (specifically, the null hypothesis). So the logic of falsification is integral to everything we will do from this point forward in this course.

A quick note about confirmation and Bayesian statistics.

Confirmation does exist.

Falsification is an ideal that many scientists strive for. And it is fundamental to frequentist statistics. However, I don't want to leave you with the idea that confirmation does not exist in modern science. It does.

Here is a famous example to show you that many people do make decisions based on confirmation:

Let's say that you are asked to design a new bridge. You have two choices:

1. An old design that has been used for hundreds of bridges, none of which have collapsed.

2. A brand new design that has never been tested before.

Which would you choose?

I'd choose #1!

Falsification says that both bridges are equal. Both have "not yet been falsified."

Confirmation says that the bridge that has been tested millions of times (each car that has driven over it) and not failed is superior.

Probabilities may allow for confirmation

The problem of confirmation teaches us that positive evidence is compatible with an infinite number of theories.

But this does not mean that the evidence is equally compatible with each theory.

Let's say you've observed 300 black ravens, and no white ravens.


x 300


x 0

This observation is compatible with an infinite number of theories:

300 ravens are black, the rest are white.

301 ravens are black, the rest are white.

302 ravens are black, the rest are white.

…

95% of ravens are black, 5% are white.

100% of ravens are black.

But these are relatively unlikely. It is unlikely that you just happened to find all of the black ravens and none of the white ones!

These are more likely.

Bayesian statistics

This intuition suggests that, even though positive evidence is compatible with an infinite number of theories, positive evidence can suggest that some theories are more likely than other theories.

So what we want to do is develop a precise way to conclude how likely a theory is given a piece of **positive** evidence.

And here is an equation that might do it for us. It is called **Bayes Theorem**.

$$P(\text{theory} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{theory}) \times P(\text{theory})}{P(\text{evidence})}$$

Don't worry. You don't need to know this slide at all for this class. We are doing **frequentist statistics**. But I want you to know how frequentist statistics compares to **Bayesian statistics.** And there are two major differences: they differ in the definition of probability (objective/subjective), and they differ in whether they pursue **falsification** or **confirmation**.

Thomas Bayes
1701-1761