Statistical theory: Probability

Research Methods for Human Inquiry

Andrew Perfors

Today's story…

Doggie and Flopsy have just returned from stealing the dataset from Otherland

Today's story…

… where LFB and Foxy have gone missing!!!

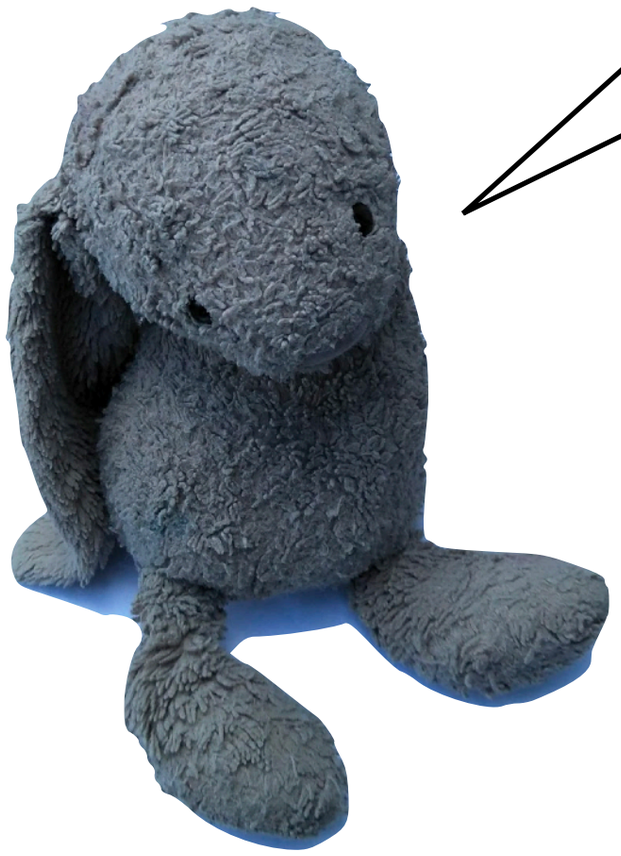
???

We need to analyse this dataset!

WHO CARES
when LFB and Foxy
are missing?!

Their sacrifice cannot have been in vain! Also, maybe we'll learn something that will help us rescue them.

But.. how will it help us learn something? How can we conclude anything from one dataset about Otherland in general?



Ah…. this is the deep question

What are we doing when we're making inferences based on our data?

(and how are we doing it?)

Three tightly linked questions

- **Descriptive statistics**: describe your **dataset**
  - Not trying to draw a conclusion beyond it, just understand what I've seen
  - This is what we've done so far (figures, summary stats)
- **Probability theory**: make a **prediction**
  - I know the truth about how my observations are being generated
  - e.g., I know the coin is fair, and so the probability of heads is 50%
  - What I want to know is which events I should expect?
  - What's the chance of seeing 32 heads and 12 tails?
- **Inferential statistics**: draw a **conclusion**
  - I know what the data are, and I want to know how they came to be
  - e.g., I've seen 32 heads and 12 tails
  - Should I conclude that this coin is fair, or biased?

- **Descriptive statistics**: describe your **dataset**
- **Probability theory**: make a **prediction**
- **Inferential statistics**: draw a **conclusion**

We often want to draw a conclusion about the population (inferential statistics) using our dataset (descriptive statistics). To do that, we use the tools of probability theory

Yeah. We want to learn about Otherland based on our dataset that we stole!

First, we need to understand probability.

What is it?

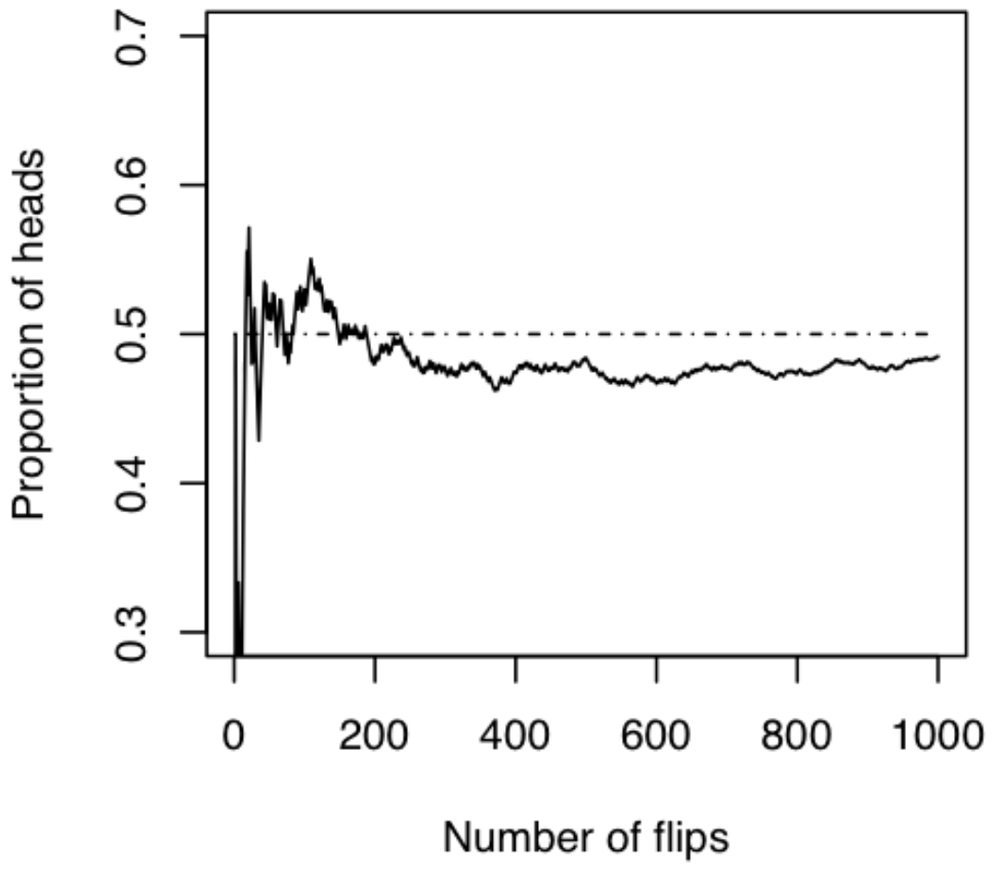
The frequentists

- Probability is “**long run frequency**”
  - If you flip a perfectly fair coin a LOT of times, about half will be heads

HTHHTTHTTHHTHTH

The frequentists

- Probability is "**long run frequency**"
  - If you flip a perfectly fair coin a LOT of times, about half will be heads
  - For *any* finite number of observations, it won't be exactly 50%
  - But if you flipped an "infinite" number of coins, it would be



In the long run, the proportion of heads ought to end up "converging" to a value of 0.5

For a frequentist

the coin has a probability of 50% heads	=	the long-run proportion of heads is 50%

The frequentists

For a frequentist

the coin has a probability of 50% heads	=	the long-run proportion of heads is 50%

The Bayesians

- Probability is **“degree of belief”**
  - What does an idealised, rational agent believe will happen?
  - But how does that play out in practice?

- Operationalised in terms of betting: what will you gamble on?
  - Some bets I expect to win…
    - You bet me $1 that Collingwood will win the next grand final
    - I believe they have a 5% chance of winning, so I’ll take the bet
  - Some bets I expect to lose…
    - You bet me $1 that my child will ask me for help with something trivial
    - I think the chances of this are very high, so I won’t take that bet

The Bayesians

- Probability is **“degree of belief”**
  - What does an idealised, rational agent believe will happen?
  - But how does that play out in practice?

- Operationalised in terms of betting: what will you gamble on?
  - My degree of belief in something is thus how likely I am to take a bet on it

For a Bayesian

the coin has a probability of 50% heads	=	I have a 50% degree of belief that the coin is heads

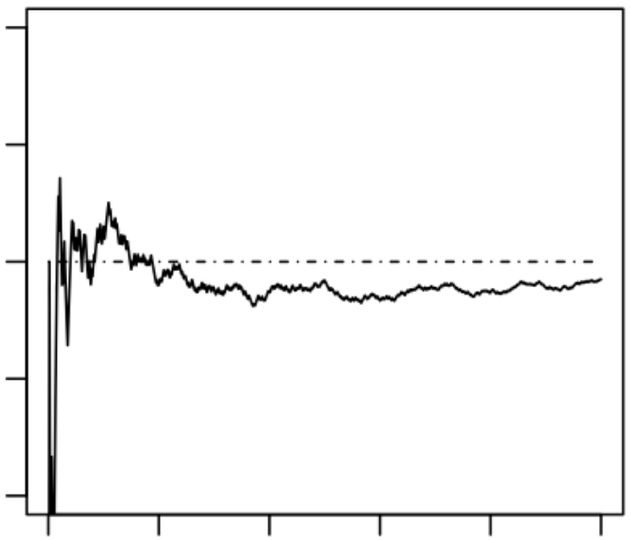
Note: there is a lot of actual research about how people actually make bets, and in practice it involves a lot more than degree of beliefs (e.g., how much money you have, how much of a risk-taker you are, etc). This formulation is a *theoretical formulation* about what probability might mean: it’s not a psychological study.

I am a Bayesian. This is the way we all intuitively think about probability anyway.

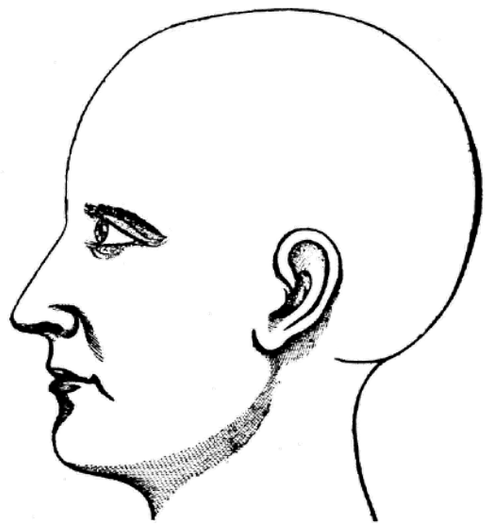
For a Bayesian

the coin has a probability of 50% heads	=	I have a 50% degree of belief that the coin is heads

Frequentists



- Probability is objective
- It is in the world
- It applies only to repeatable events



Bayesians

- Probability is subjective (but not arbitrary)
- It is in the beliefs of a rational agents
- It applies to anything you can believe in

Does it really matter?

- In one sense, no
  - People intuitively switch between both versions
  - “I think Melbourne Storm will win the next grand final” (Bayesian)
  - “1 in 3 people will get cancer” (frequentist)
- In a bigger sense, yes
  - What is the probability that the Higgs boson exists?
  - Bayesian: in light of current evidence, I think it’s 91% (or whatever)
  - frequentist: the “existence of Higgs boson” is not a repeatable event.
  - This difference changes the way you go about doing statistics!

Who is right?

- Eek! That's a hard question
  - I use Bayesian tools sometimes and frequentist tools at other times
  - I think Bayesian methods are more powerful, but they're trickier
- The history:
  - There was a big fight over this in the early 20th century
  - Frequentists won the fight in statistics… but only temporarily
  - Bayesian probability has made a big comeback
  - The old fight is currently being re-fought in psychology
- The status quo:
  - Most of the tools I'll teach here are frequentist. Because that's the "norm"

So… what do we know?

Probability is just a number

The probability of *everything that can happen* must sum to 1

Example: you flip a coin two times. What are the probabilities?

T T	0.25
H T	0.25
T H	0.25
H H	0.25

← This is okay

0

will certainly not happen

1

will certainly happen

Probability is just a number

The probability of *everything that can happen* must sum to 1

Example: you flip a coin two times. What are the probabilities?

T T	0.4
H T	0.5
T H	0.2
H H	0.3

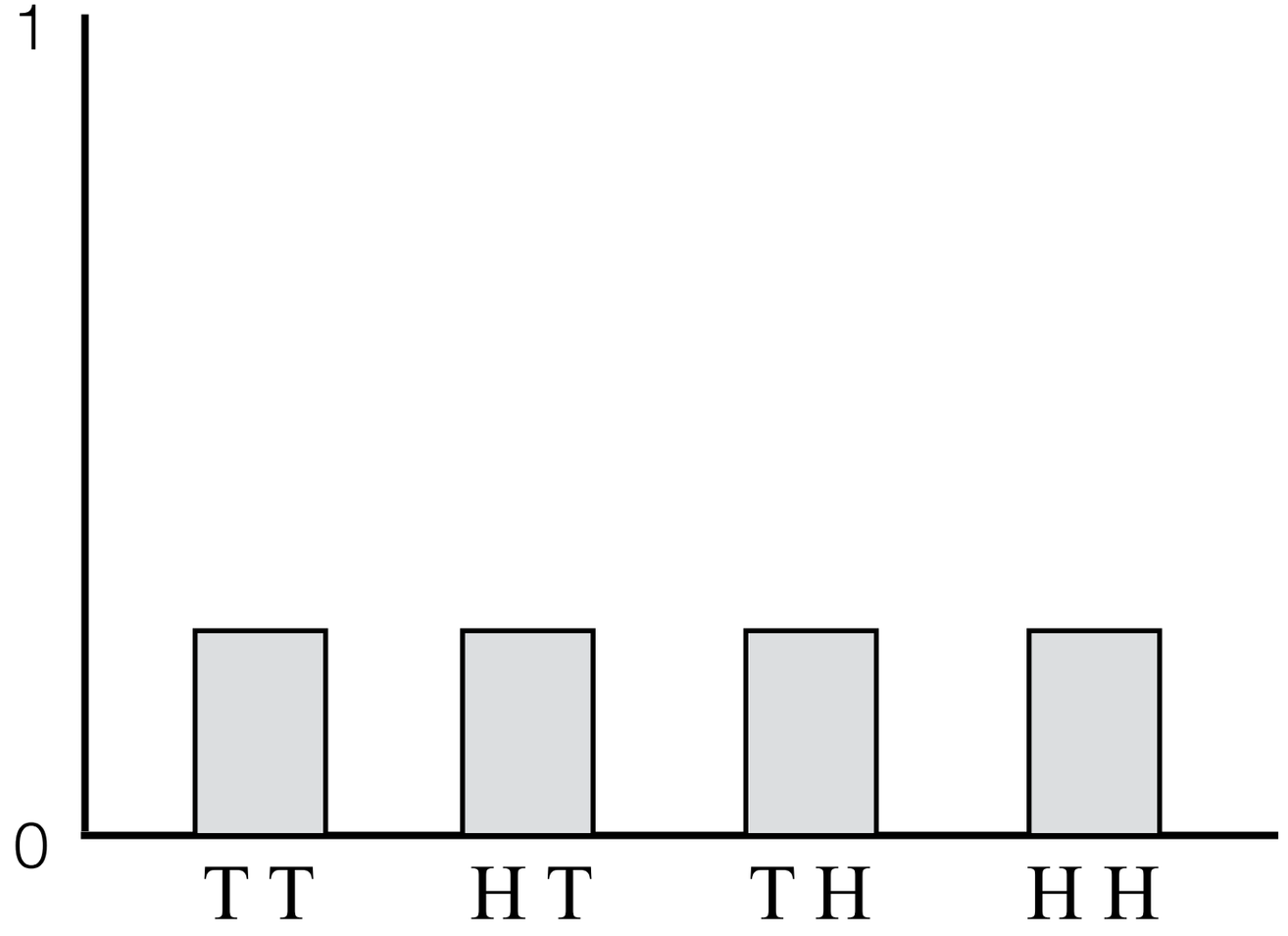
← This is NOT

0 — will certainly not happen

1 — will certainly happen

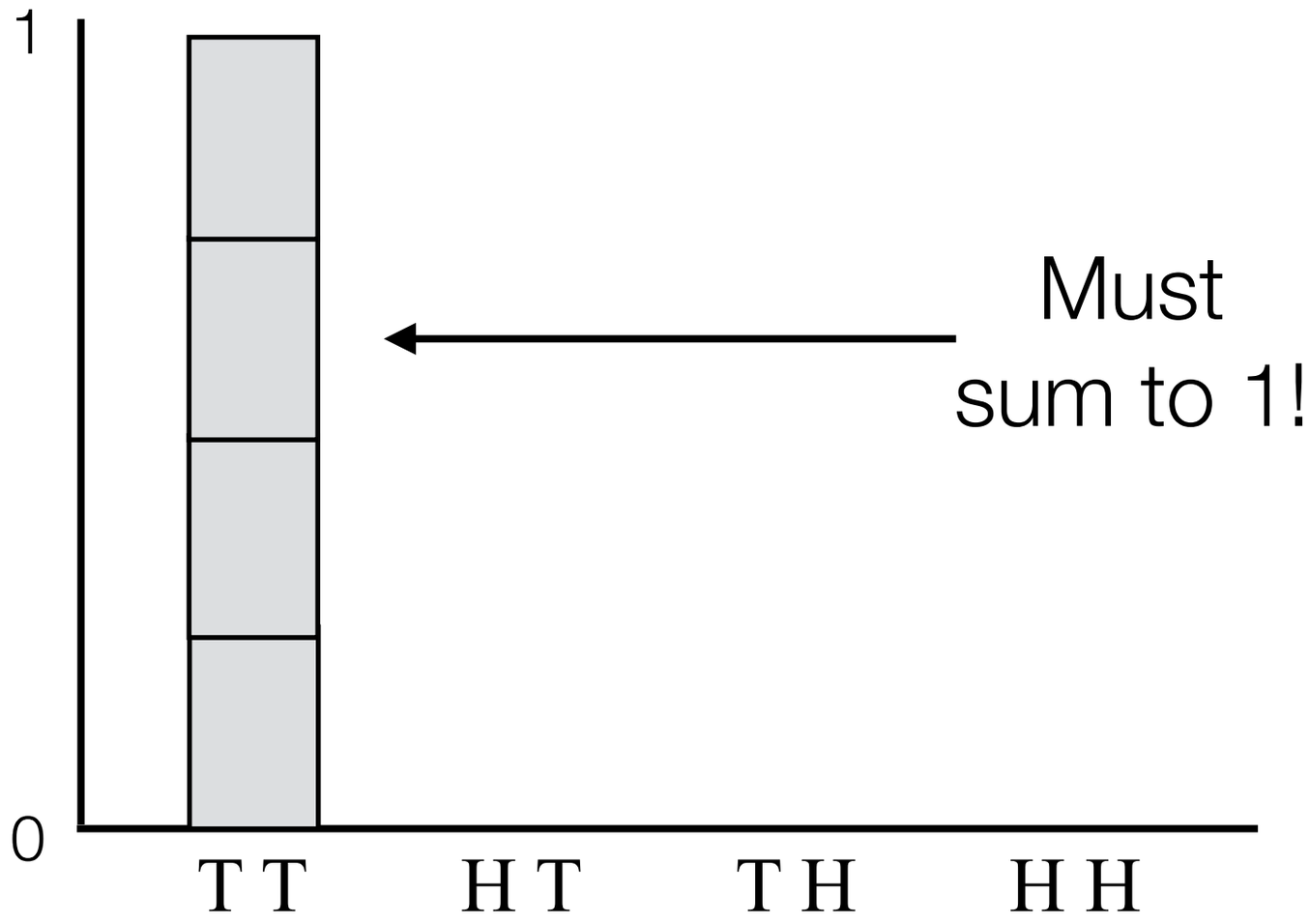
Probability is just a number

We can draw a probability distribution
by putting the probability on the y axis and the events on the x axis.

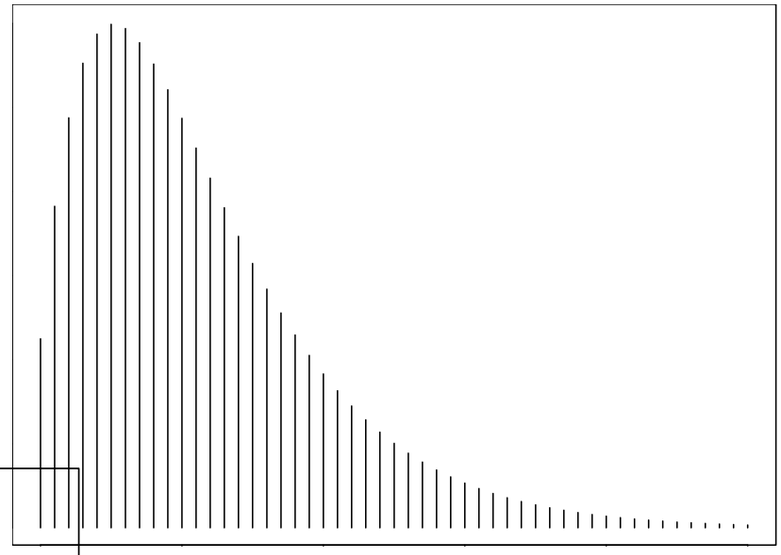


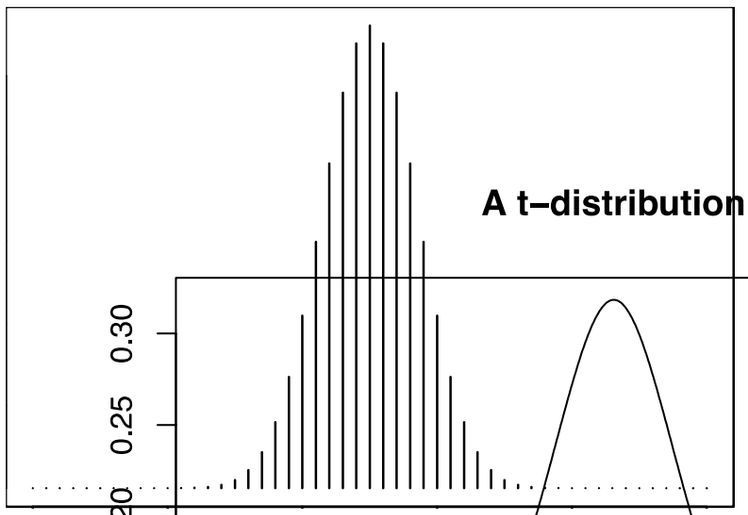
Probability is just a number

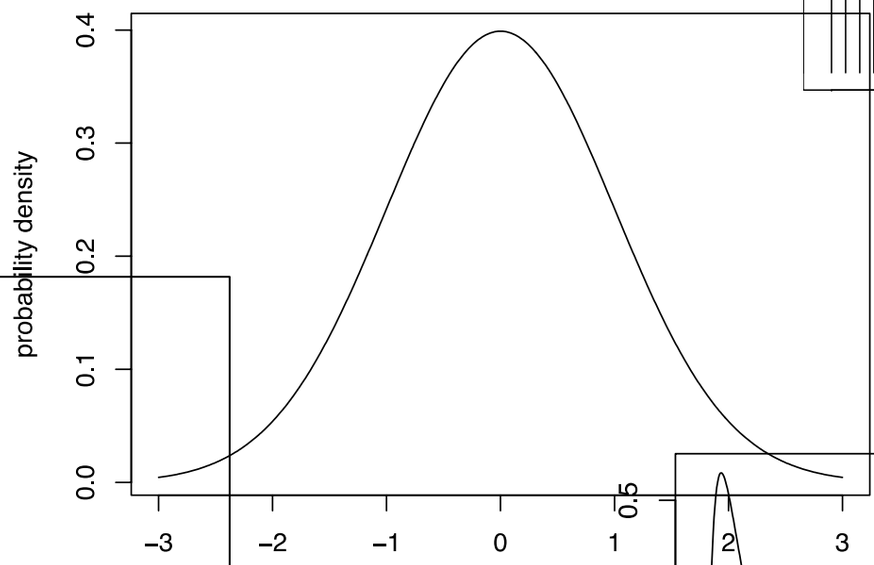
We can draw a probability distribution
by putting the probability on the y axis and the events on the x axis.

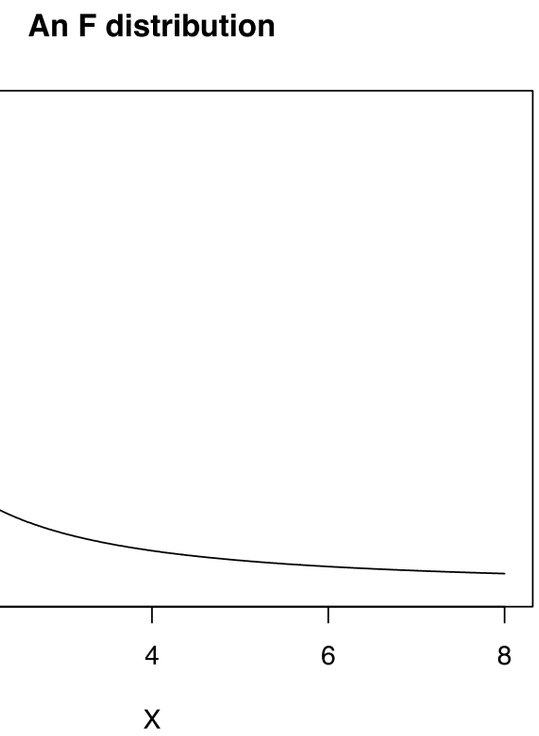


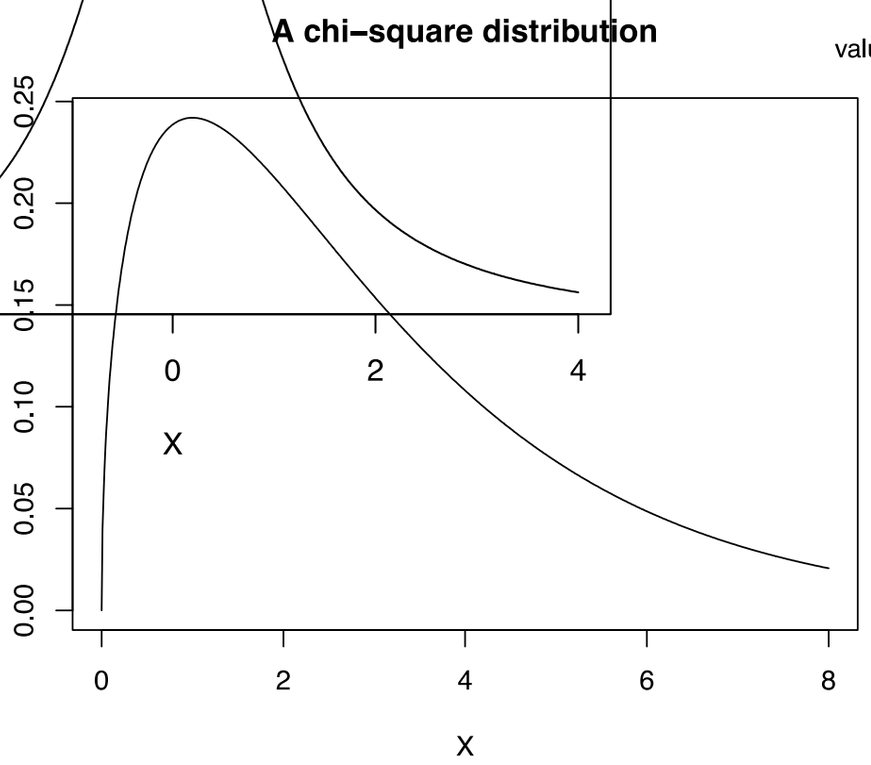
There are lots of types of probability distribution…



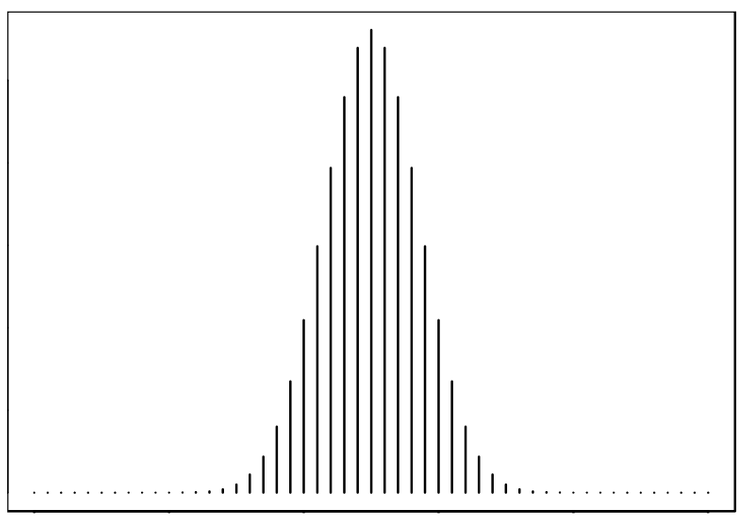




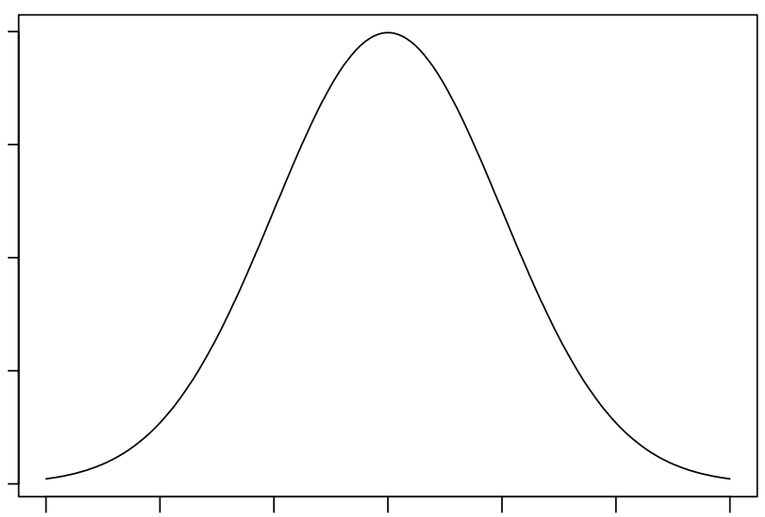




I’ll talk about two of them



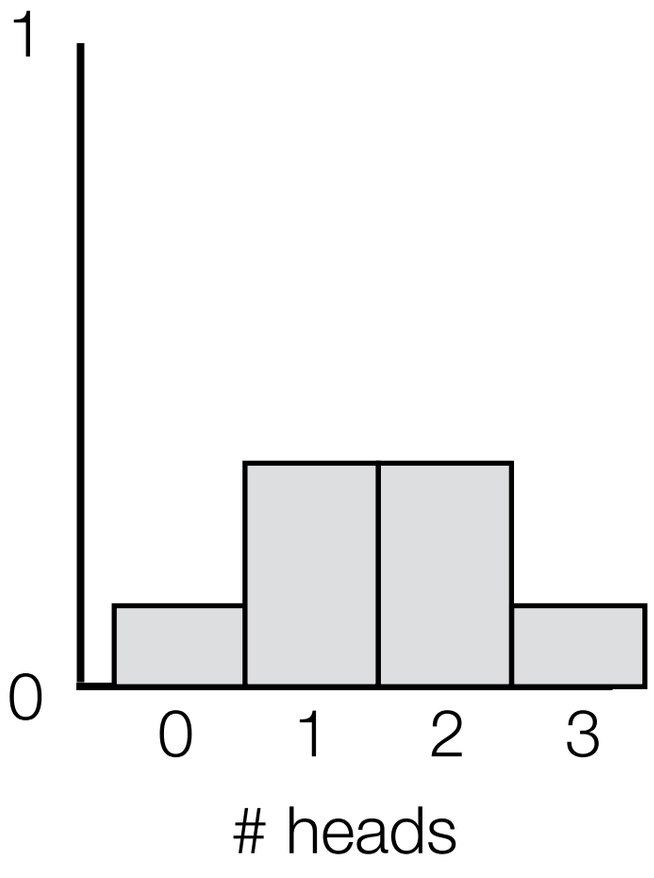
binomial



normal

Binomial distribution

- The binomial distribution is used to describe count data of one of two possible events happening
- e.g., I flip a fair coin 3 times.
  - Assume that the probability of each head is 0.5
  - What is the probability that I get 2 heads?
  - Or 1 head? Or whatever?



H H H	3 heads	1/8 = 12.5%
T H H H T H H H T	2 heads	3/8 = 37.5%
T T H T H T H T T	1 head	3/8 = 37.5%
T T T	0 heads	1/8 = 12.5%

8 total possibilities

Binomial distribution

- The binomial distribution is used to describe count data of one of two possible events happening
- e.g., I flip a fair coin 3 times.
  - Assume that the probability of each head is 0.5
  - What is the probability that I get 2 heads?
  - Or 1 head? Or whatever?
- There's an equation that tells you the answer:
- But you do not have to know it!! Erase it from your memory!
  - Let's use R instead.

Working with probability distributions

- R has a lot of functionality to let you play with distributions
- It's always the same structure:
  - `dbinom()` - Probability (density) of a specific outcome
  - `pbinom()` - Chance that the outcome doesn't exceed a threshold
  - `qbinom()` - Compute some quantile of the distribution
  - `rbinom()` - Sample a random number from a distribution

Working with probability distributions

- R has a lot of functionality to let you play with distributions
- It's always the same structure:
  - `dbinom()` - Probability (~~density~~) of a specific outcome

```
> dbinom( x=21, size=50, prob=0.5 )
```

calculates the probability of a specific outcome given a binomial distribution

number of "successes" (which we'll define as heads)

total number of trials (i.e., coin flips)

the "weighting" of the coin flip assumed (0.5 is unbiased)

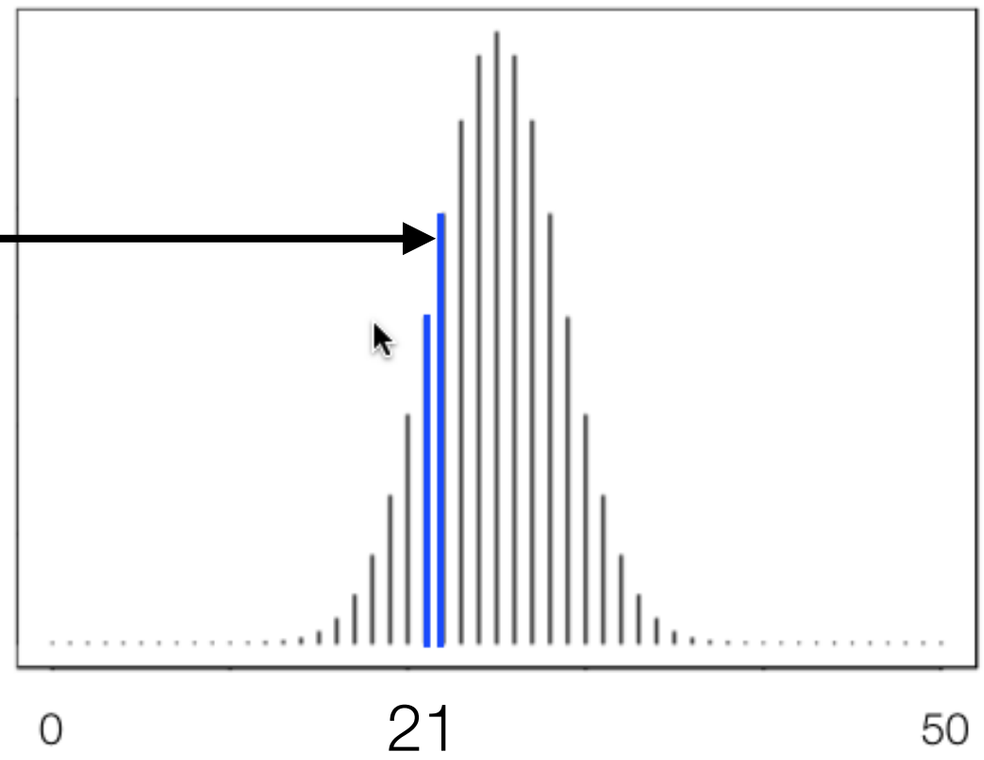
Working with probability distributions

```
> dbinom( x=21, size=50, prob=0.5 )
[1] 0.05979878

> dbinom( x=22, size=50, prob=0.5 )
[1] 0.07882567
```

5.9% probability of getting 21 heads if you flipped a coin 50 times!

7.9% of 22 heads with 50 flips!

returns this value (the height of this bar)

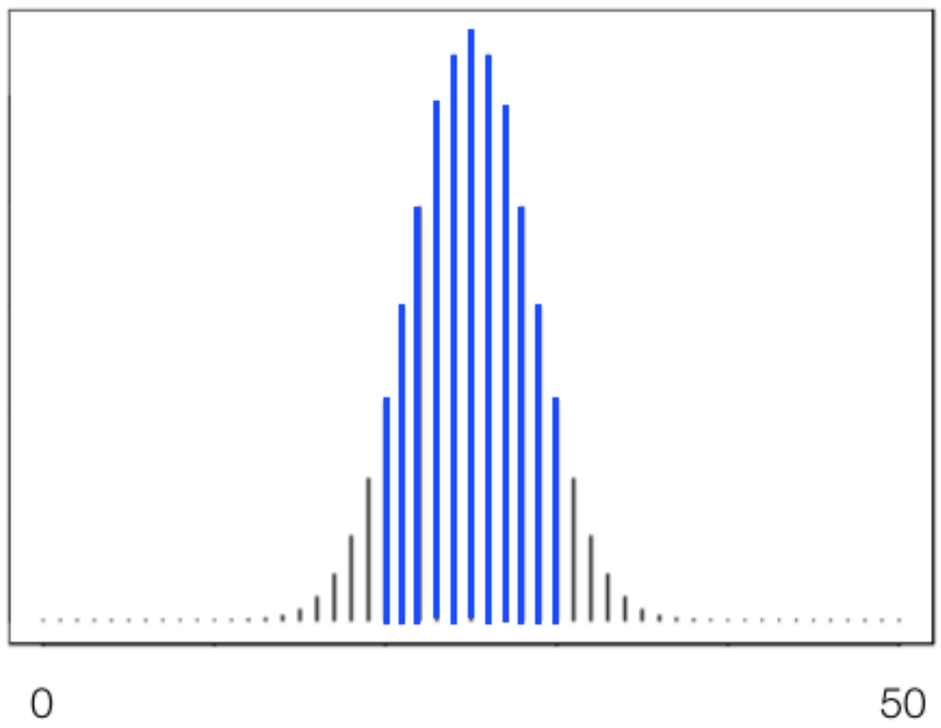


13.8% probability of either 21 or 22 heads

Working with probability distributions

How about the probability of getting between 20 and 30 heads?

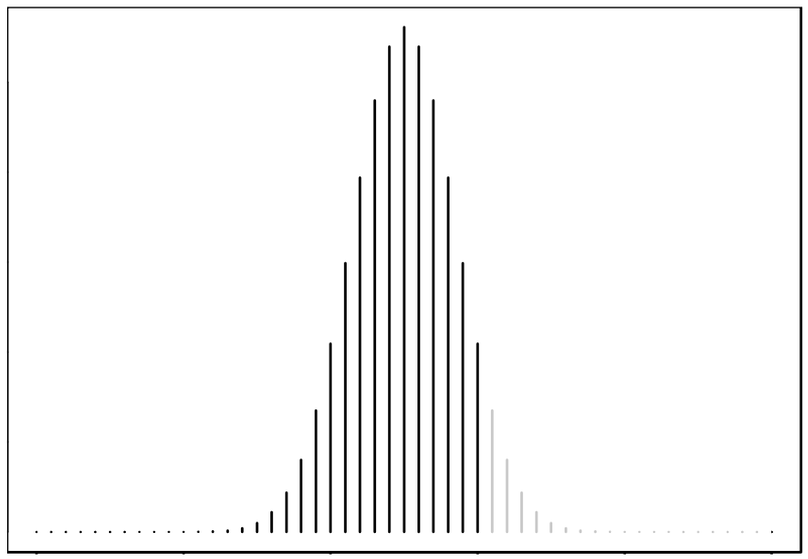
```
> dbinom(x=20:30, size=50, prob=0.5)
 [1] 0.04185915 0.05979878 0.07882567 0.09596169 0.10795690 0.11227517
 [7] 0.10795690 0.09596169 0.07882567 0.05979878 0.04185915

> sum( dbinom(x=20:30, size=50, prob=0.5) )
[1] 0.8810795
```



The probability of getting somewhere between 20 and 30 heads is 88%

Another way of getting the same thing

`pbinom()` Chance that the outcome doesn't exceed a threshold



The probability of getting 30 heads or fewer is **??**

```
> pbinom( q=30, size=50, prob=0.5 )
```

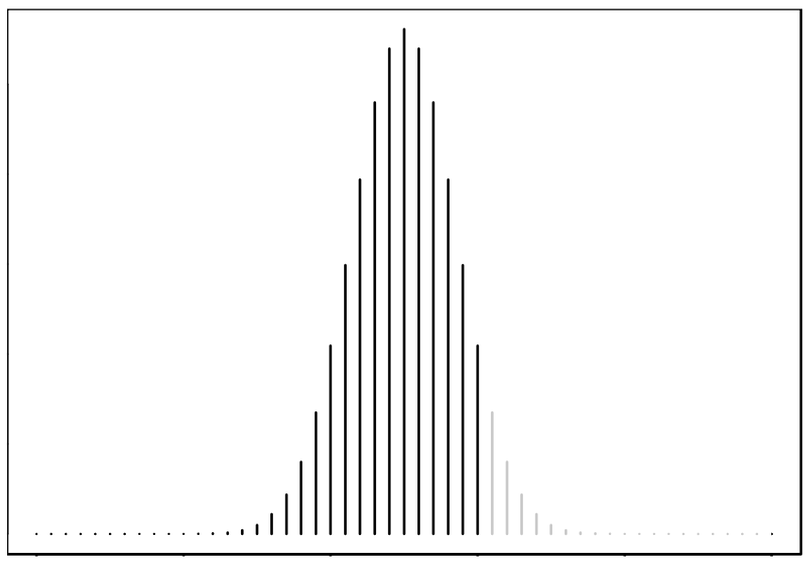
threshold

trials

bias of coin

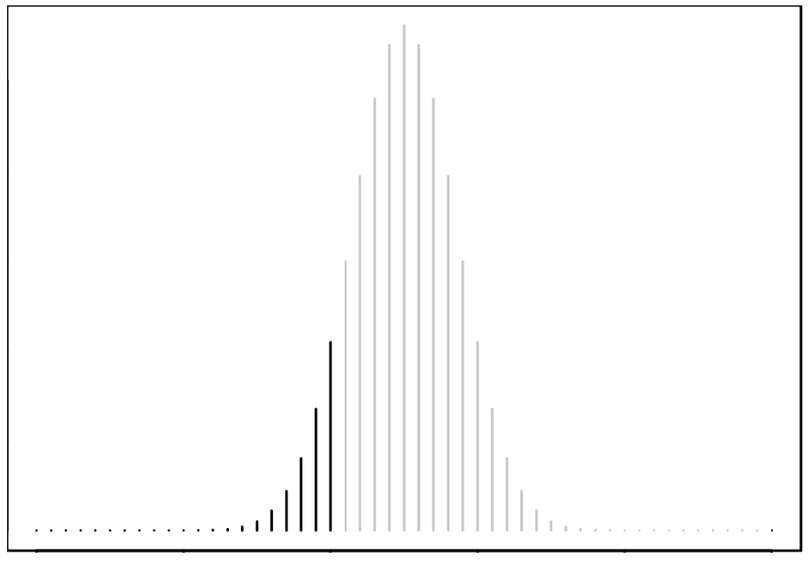
Another way of getting the same thing

`pbinom()` Chance that the outcome doesn't exceed a threshold



The probability of getting 30 heads or fewer is **94%**

```
> pbinom( q=30, size=50, prob=0.5 )
[1] 0.9405398
```



The probability of getting 19 heads or fewer is **6%**

```
> pbinom( q=19, size=50, prob=0.5 )
[1] 0.05946023
```

Other things you can do

`qbinom()` Calculate the quantile of the distribution

```
> qbinom(p=0.5,size=50,prob=0.5)
[1] 25
```

- `p=0.5` — quantile (if you put 0.5 that is the median)
- `size=50` — # trials
- `prob=0.5` — bias of coin

if you flip a coin 50 times the median number of heads you'll see is 25

Other things you can do

`qbinom()` Calculate the quantile of the distribution

```
> qbinom(p=c(0.1,0.5,0.9),size=50,prob=0.5)
[1] 20 25 30
```

the 90% percentile is 30

if you flip a coin 50 times the median number of heads you'll see is 25

the 10% percentile is 20

Other things you can do

`rbinom()` Sample random numbers from the distribution

```
> rbinom(n=10,size=50,prob=0.5)
[1] 27 22 22 31 25 22 23 24 23 26
```

first set of 50, got 27 heads

second set, got 22 heads

number of sets to sample

** note that since this is random, you will get different numbers than I did!*

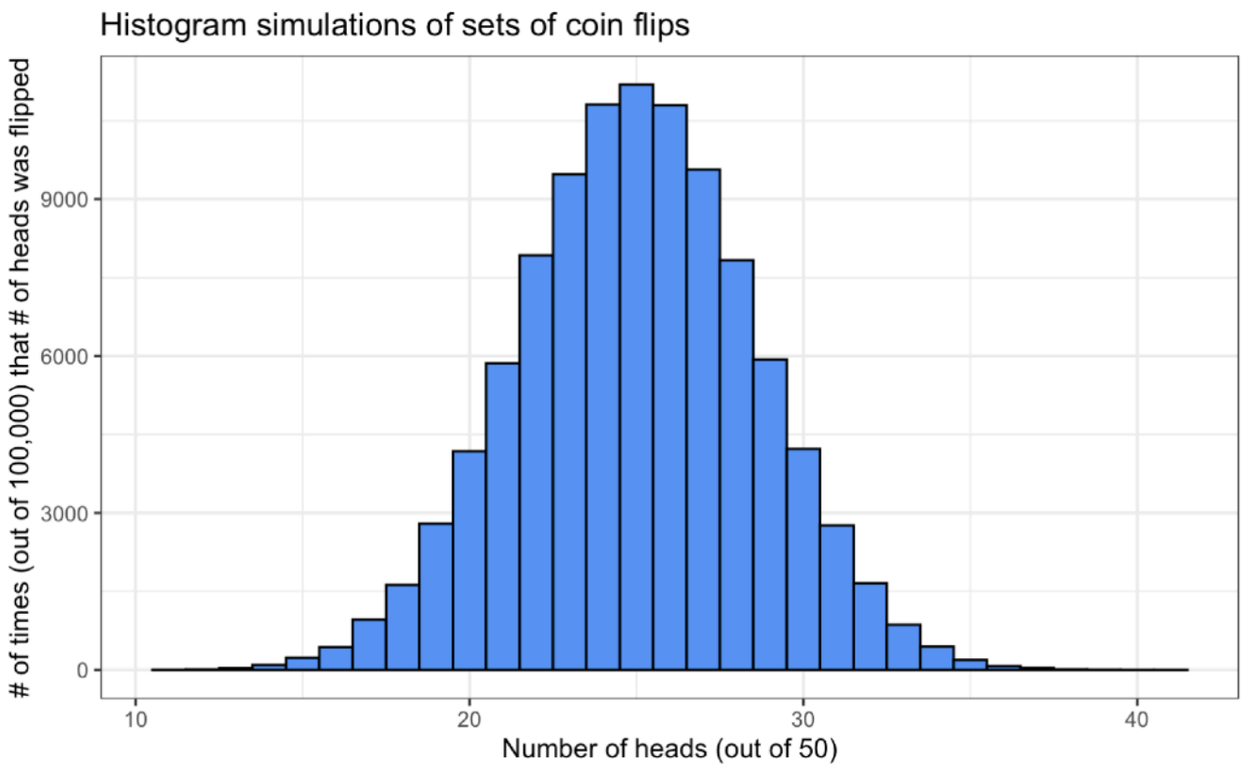
Other things you can do

`rbinom()` Sample random numbers from the distribution

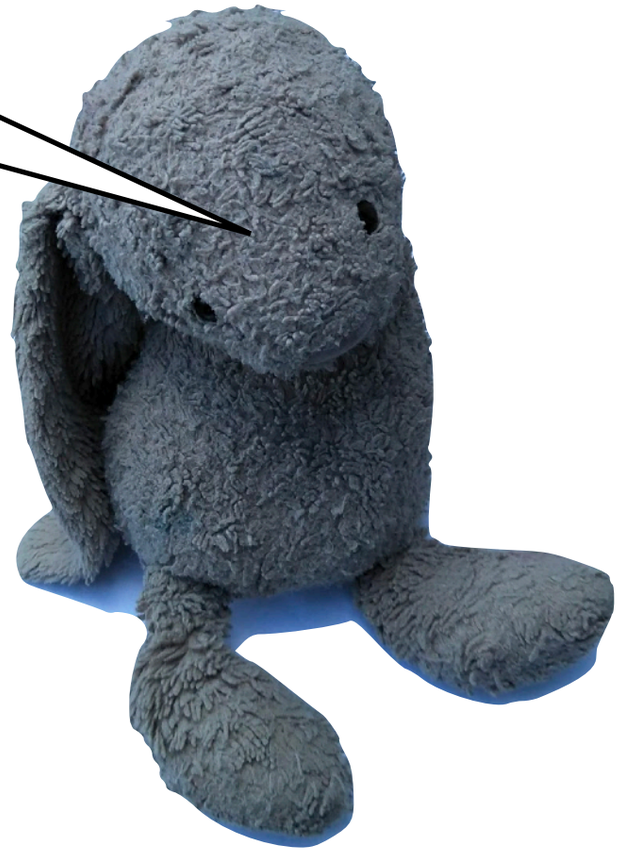
```
> rbinom(n=10,size=50,prob=0.5)
[1] 27 22 22 31 25 22 23 24 23 26
```

We can generate a good binomial histogram by generating thousands of sets!

```
> binomialData <- rbinom(100000,size=50,prob=0.5)
```

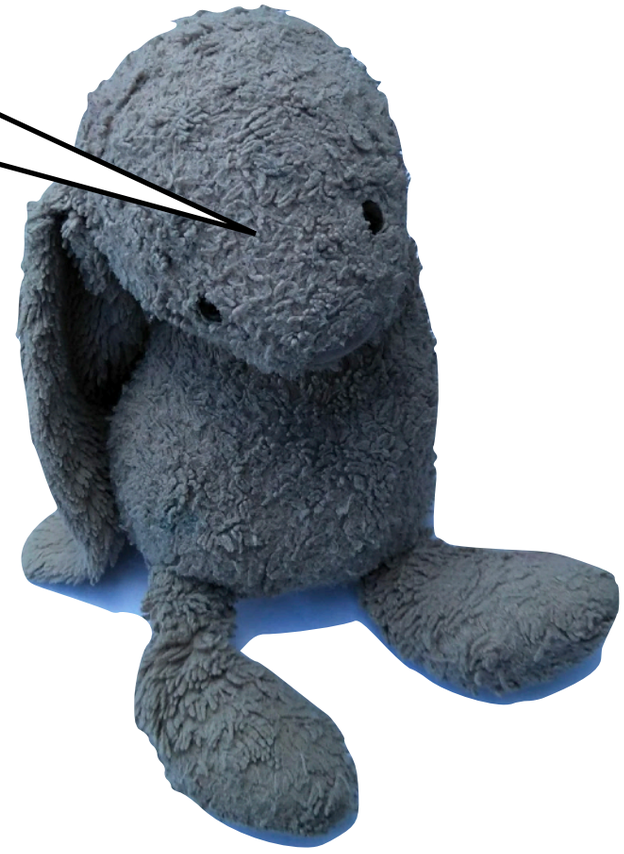


But… why should I care about coin flips? That has nothing to do with psychology or our problems or *anything*

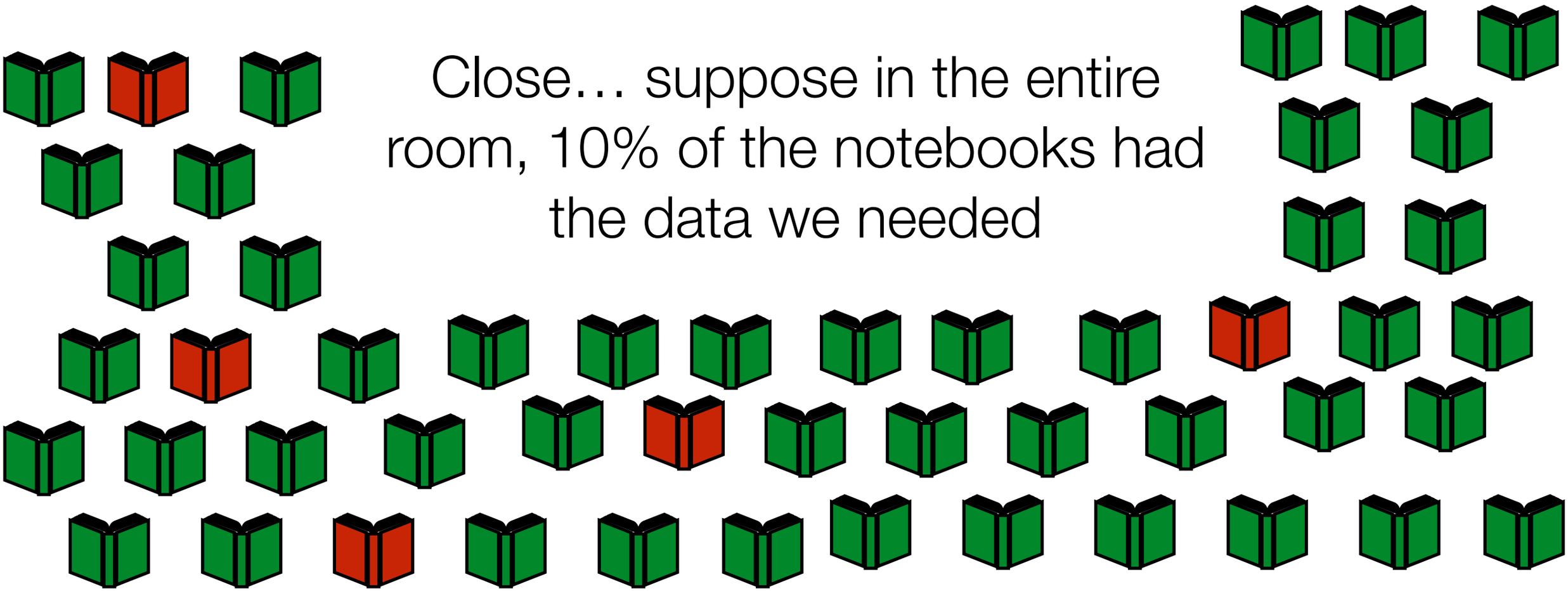


The binomial distribution can capture *anything* where there are two possible outcomes and there is some underlying probability of success.

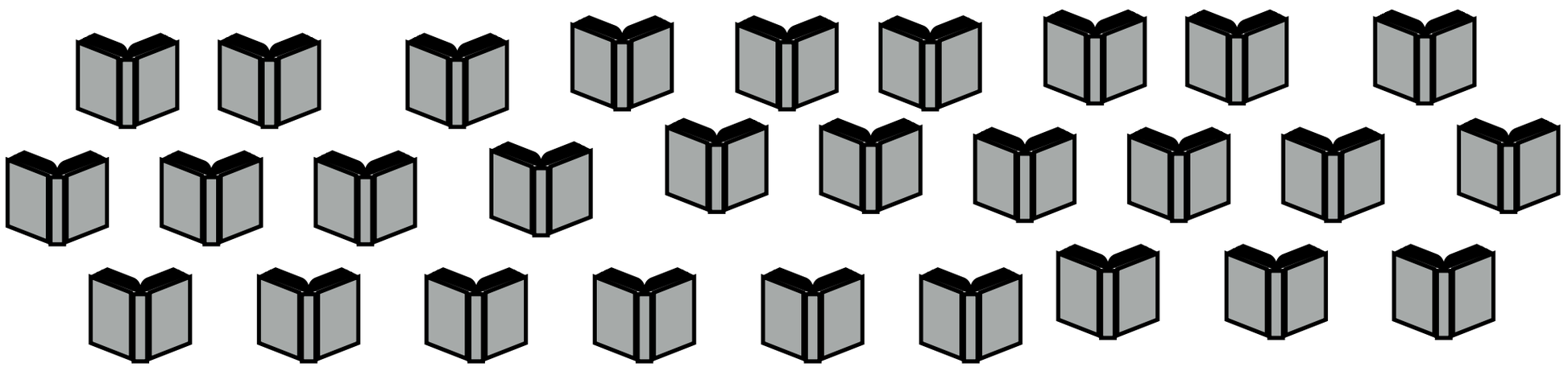
Hmmm… like the probability that the 20 notebooks Flopsy grabbed would have the data we needed?



Close… suppose in the entire room, 10% of the notebooks had the data we needed



Calculations can tell us what we might expect to see in the 20 notebooks she grabbed



20 notebooks: what can we expect?

(Assume that in general the underlying probability is 10% - it's hard)

```
> dbinom(x=3, size=20, prob=0.1)
[1] 0.1901199
```

Probability that exactly 3 notebooks contain that data

```
> pbinom(q=3, size=20, prob=0.1)
[1] 0.8670467
```

Probability that 3 or fewer notebooks contain that data

```
> qbinom(p=0.5,size=20,prob=0.1)
[1] 2
```

Median number of notebooks we might expect to contain the data

```
> rbinom(n=10,size=20,prob=0.1)
[1] 1 4 3 1 5 2 0 2 2 0
```

For 10 sets of 20 each, # of notebooks with the data in each set

Getting a question right on an MCQ (10 questions)

(Assume that in general you're pretty smart, and should get around 70%)

```
> dbinom(x=10, size=10, prob=0.7)
[1] 0.02824752
```

Probability of getting a perfect score

```
> pbinom(q=5, size=10, prob=0.7)
[1] 0.1502683
```

Probability of getting 5 or less right

```
> qbinom(p=c(0.25,0.75),size=10,prob=0.7)
[1] 6 8
```

Interquartile range of what you expect to get

```
> rbinom(n=5,size=10,prob=0.7)
[1] 9 7 5 5 7
```

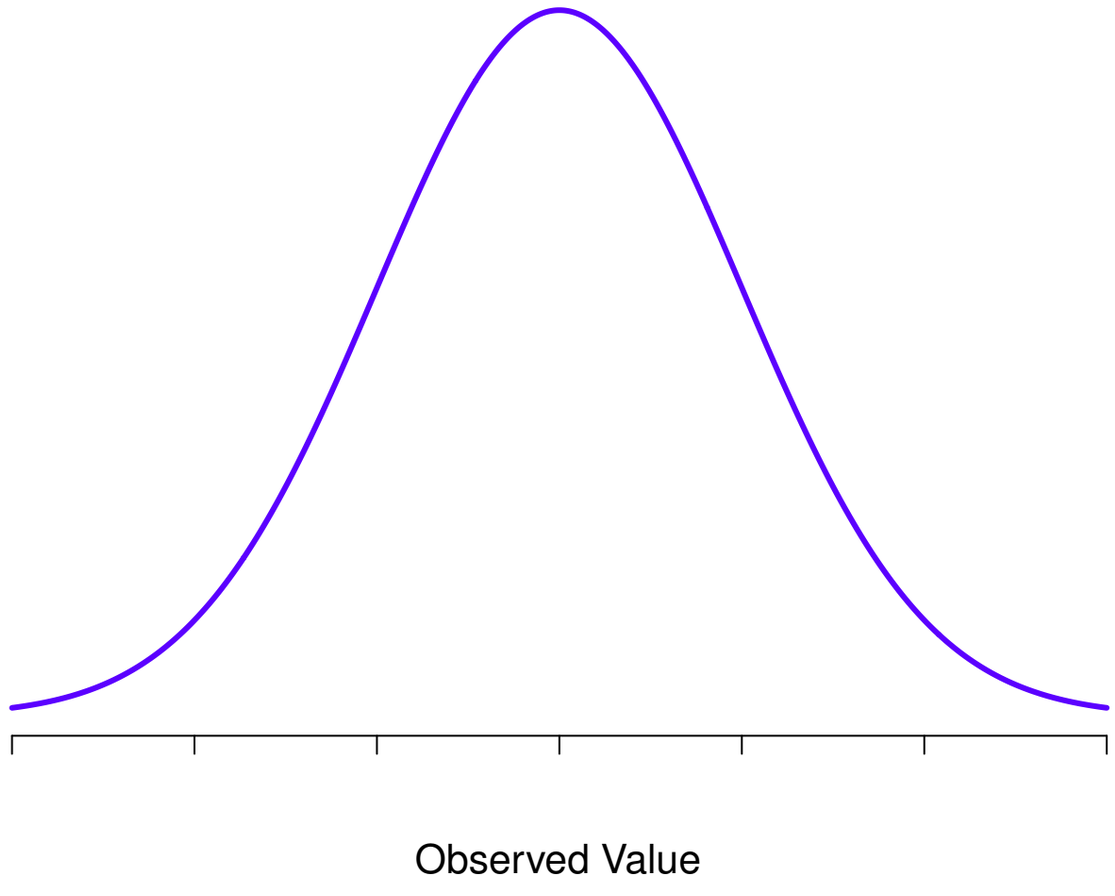
Score on each of five random quizzes

The normal distribution

(a.k.a. Gaussian, bell curve, etc)

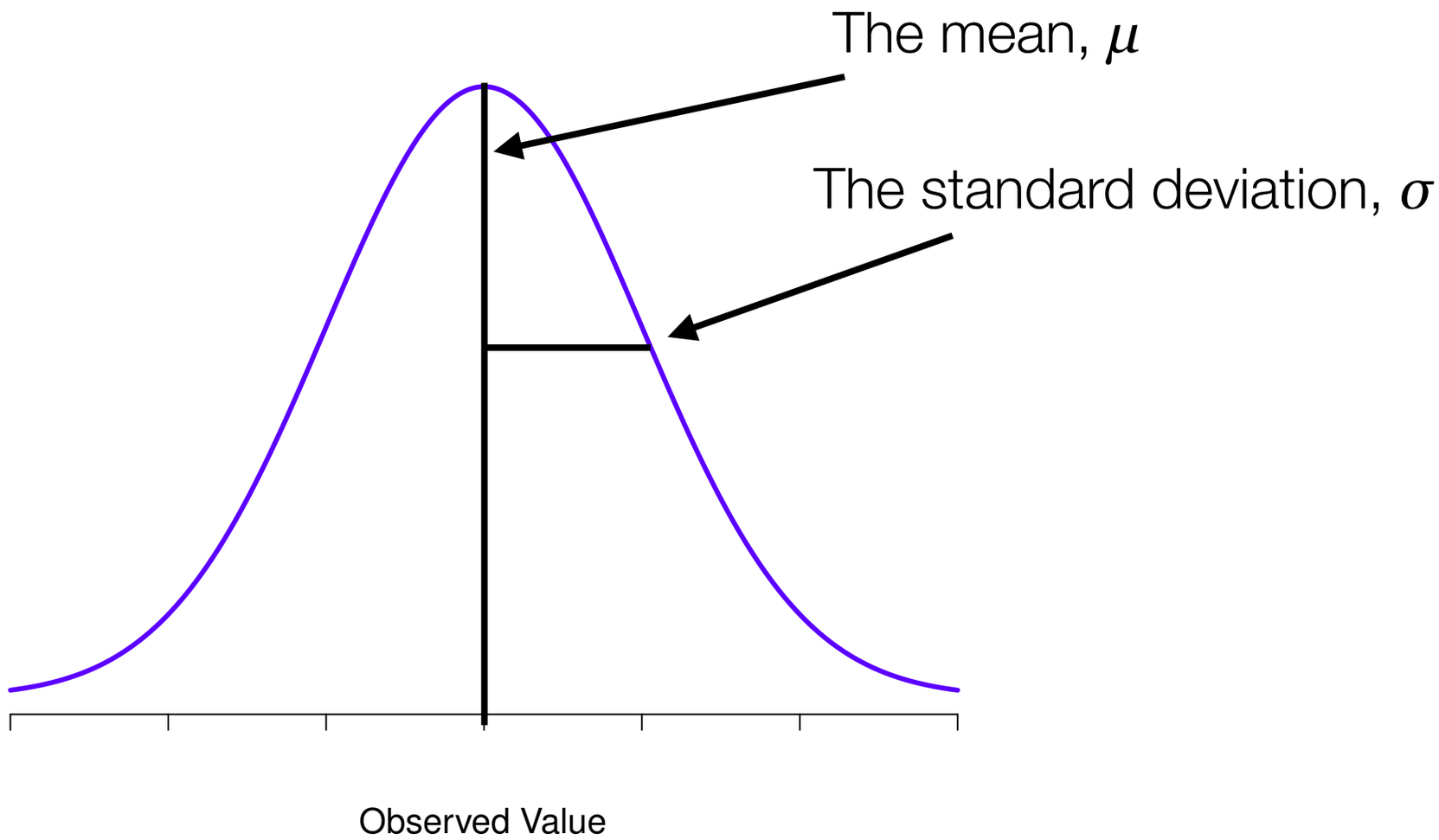
The normal distribution

- Normal distributions turn up *everywhere*
- It's very very common to see this "bell curve" shape
- I'll explain why this pattern is so common later



The normal distribution

- A normal distribution is perfectly described by two **parameters**
- The mean, μ (mu), and the standard deviation σ (sigma)
- The mean, median, and mode are also identical



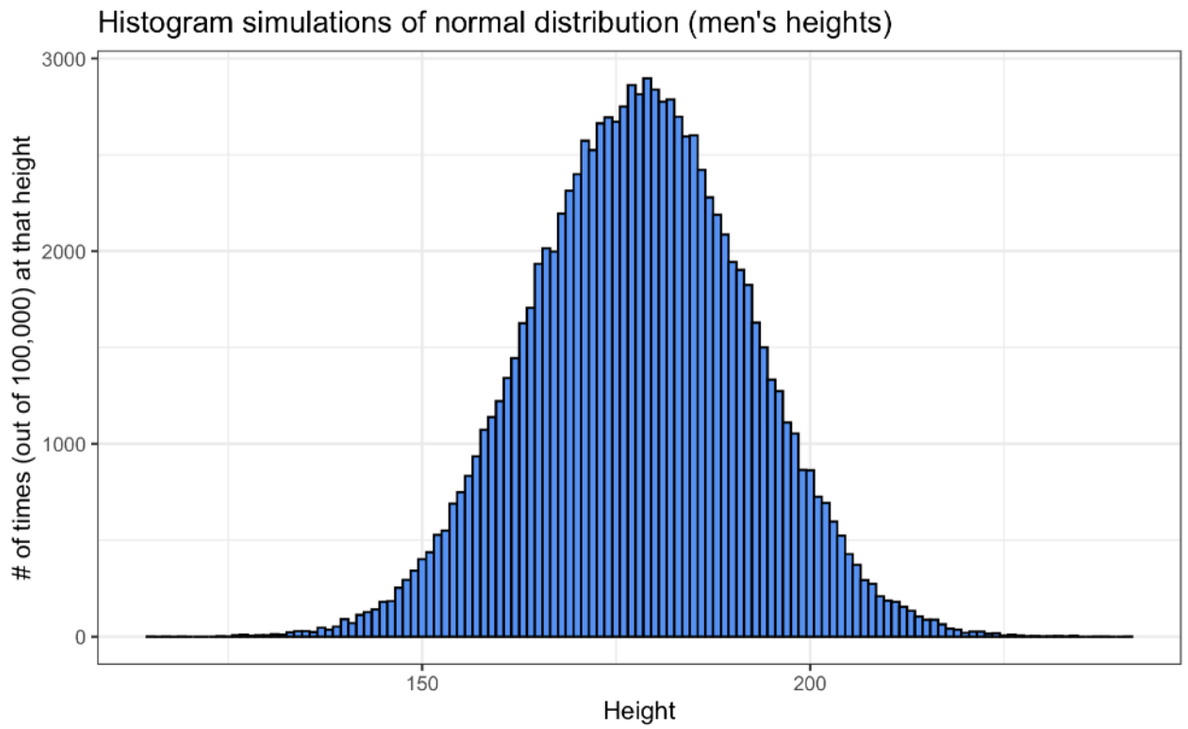
The normal distribution

- A normal distribution is perfectly described by two **parameters**
- The mean, μ (mu), and the standard deviation σ (sigma)
- The mean, median, and mode are also identical

- You can use the same kinds of functions we already saw
  - `dnorm()` - Probability (density) of a specific outcome
  - `pnorm()` - Chance that the outcome doesn't exceed a threshold
  - `qnorm()` - Compute some quantile of the distribution
  - `rnorm()` - Sample a random number from a distribution

The normal distribution

Men's heights are distributed normally around a mean of 178cm with a standard deviation of 14cm



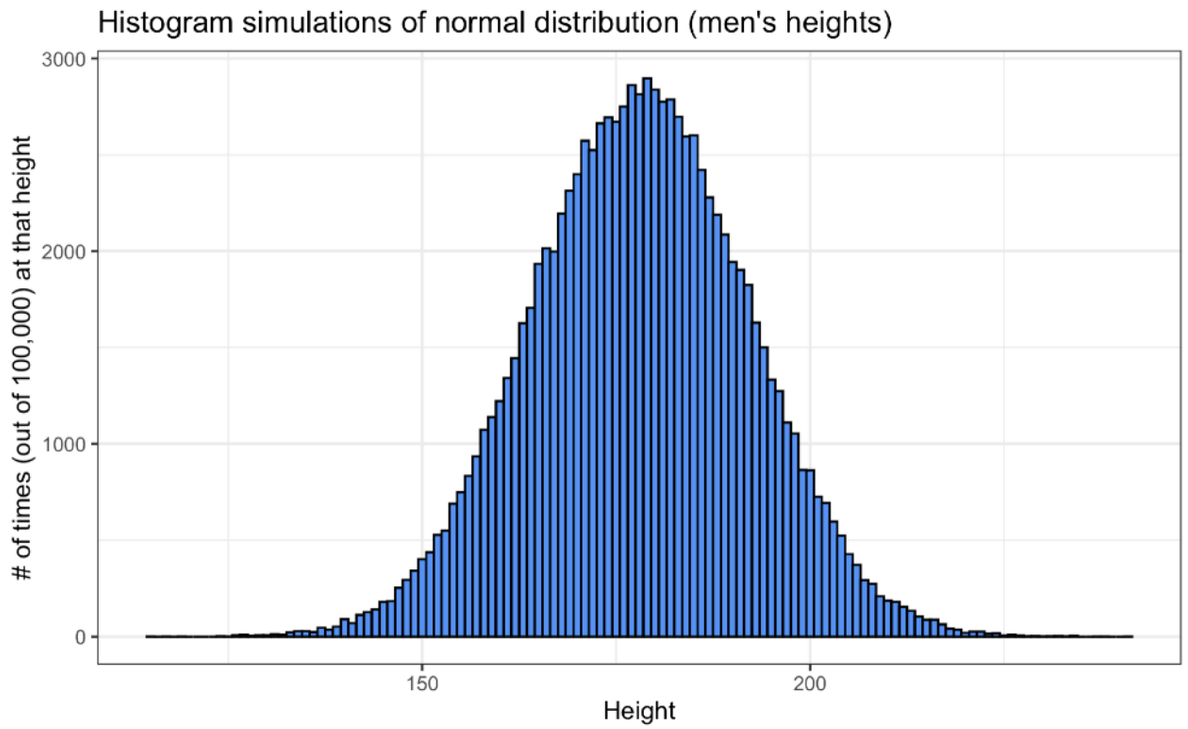
Imagine you run into 10 random guys. How tall are they?

```
> round( rnorm(n=10,mean=178,sd=14) )
[1] 185 176 198 201 180 180 205 191 173 166
```

** note that since this is random, you will get different numbers than I did!*

The normal distribution

Men's heights are distributed normally around a mean of 178cm with a standard deviation of 14cm

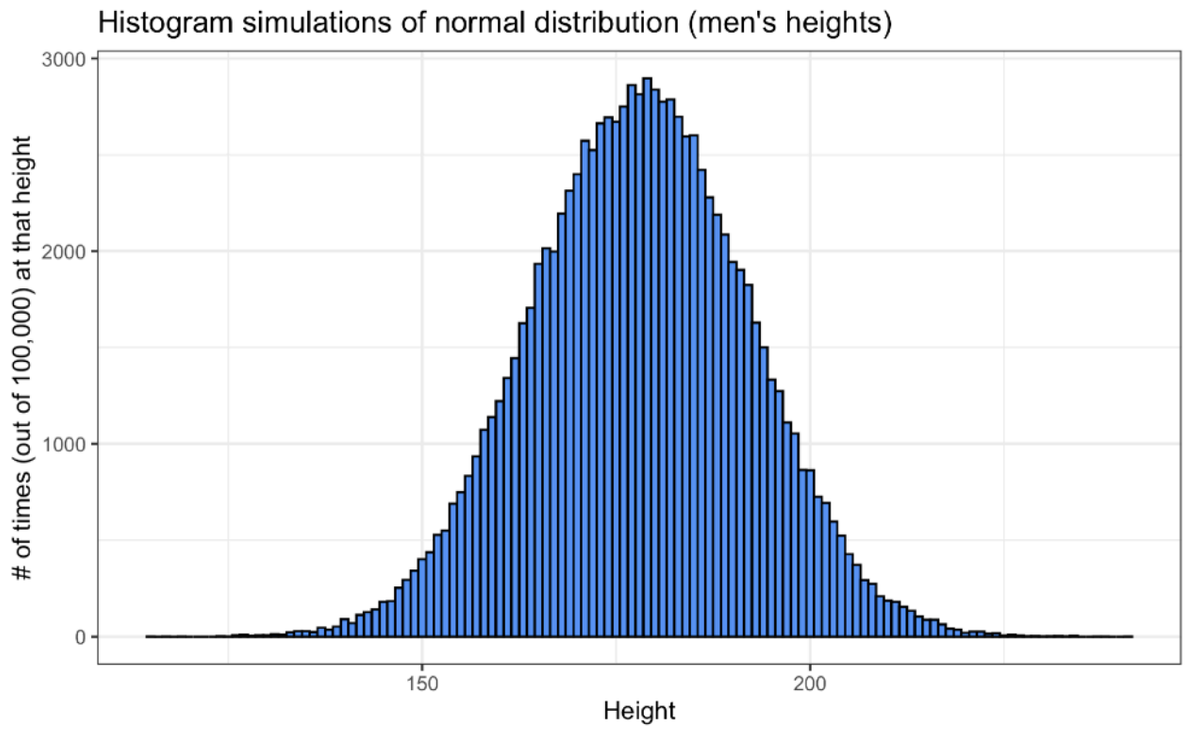


What is the median men's height?

```
> qnorm(p=0.5,mean=178,sd=14)
[1] 178
```

The normal distribution

Men's heights are distributed normally around a mean of 178cm with a standard deviation of 14cm

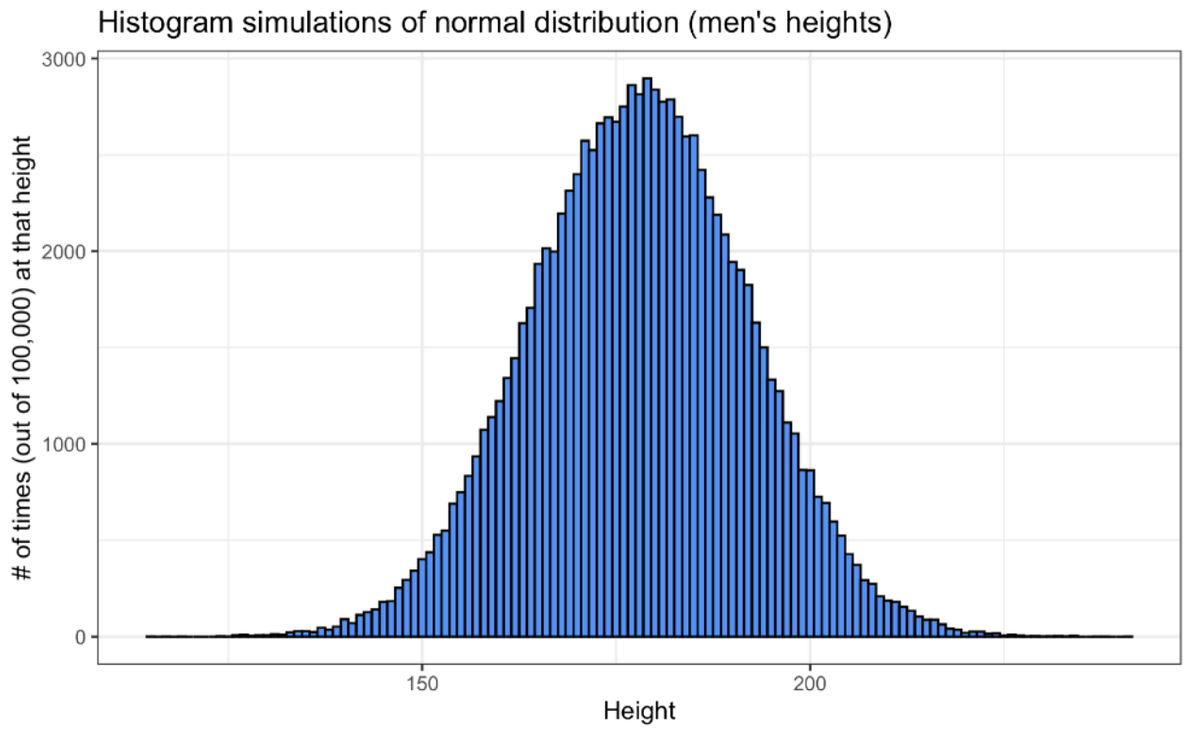


What percentage of men are equal to or shorter than 160cm?

```
> pnorm(q=160,mean=178,sd=14)
[1] 0.0992714
```

The normal distribution

Men's heights are distributed normally around a mean of 178cm with a standard deviation of 14cm



What is the probability of finding someone who is 185cm tall?

```
> dnorm(x=185,mean=178,sd=14)
```

NO!

The normal distribution

Height is a continuous variable.

Exactly 185cm?
Not 185.302 or 185.000014?
Exactly 185.00000000000000000?

The normal distribution

Height is a continuous variable.

The normal distribution applies only to continuous variables.

The probability of any specific continuous variable **makes no sense.** It is called a “probability density” and it is uninterpretable.

There is a way to deal with this but it really doesn’t matter for this class. For now, just know you should never need to use this:

`dnorm()`

See the `w5day1exercises.Rmd` file for the exercises!