Normal & Sampling Distributions

Recap

Measures of Central Tendency

- Mean (average)
- Median (middle)
- Mode (most)

Measures of Dispersion

- Variance
- Standard deviation

Standardized Scores

The Normal Distribution

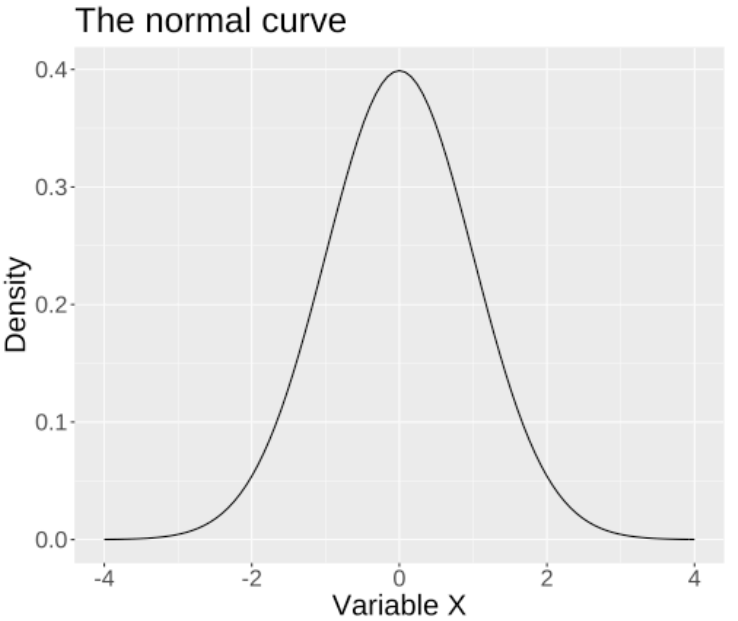
The **normal distribution**

- aka "bell curve" or "Gaussian distribution"
- Two-parameter distribution defined by the mean (μ) and standard deviation (σ)

The Normal Distribution

The **probability density function** gives the height of the curve at a particular value for X.

Although these values communicate information about probability or likelihood, they are not probabilities.



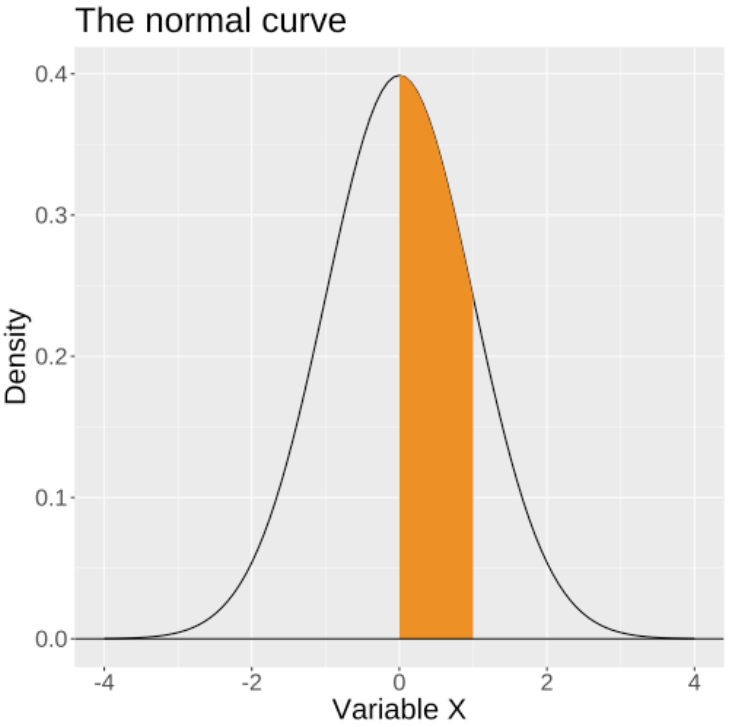
Probability Density Functions

You can take an entire semester long course on probability. Getting into the details is beyond the scope of this class, sadly. What you should know:

- The total area under the curve of a proability density function is **1**
- For a given continuous random variable, the probability of getting any single value is basically **0**

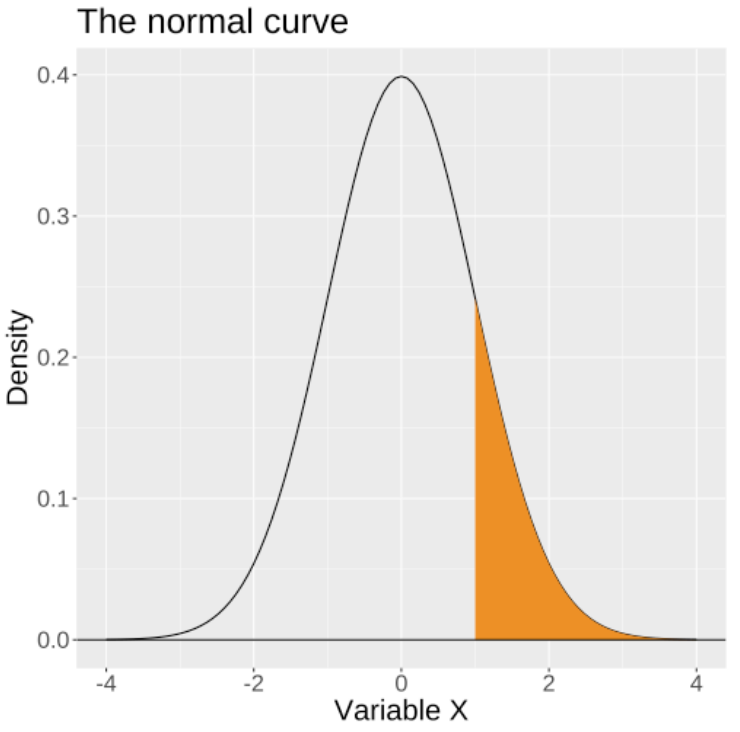
Probability of a single value

The area under the curve that lies between the mean (here 0) and a value of 1 is the probability of a score between 0 and 1.



Probability of a single value or more extreme

We can also make statements about probability like *"the probability of obtaining an x value of 1* ***or larger*** *is ___"*

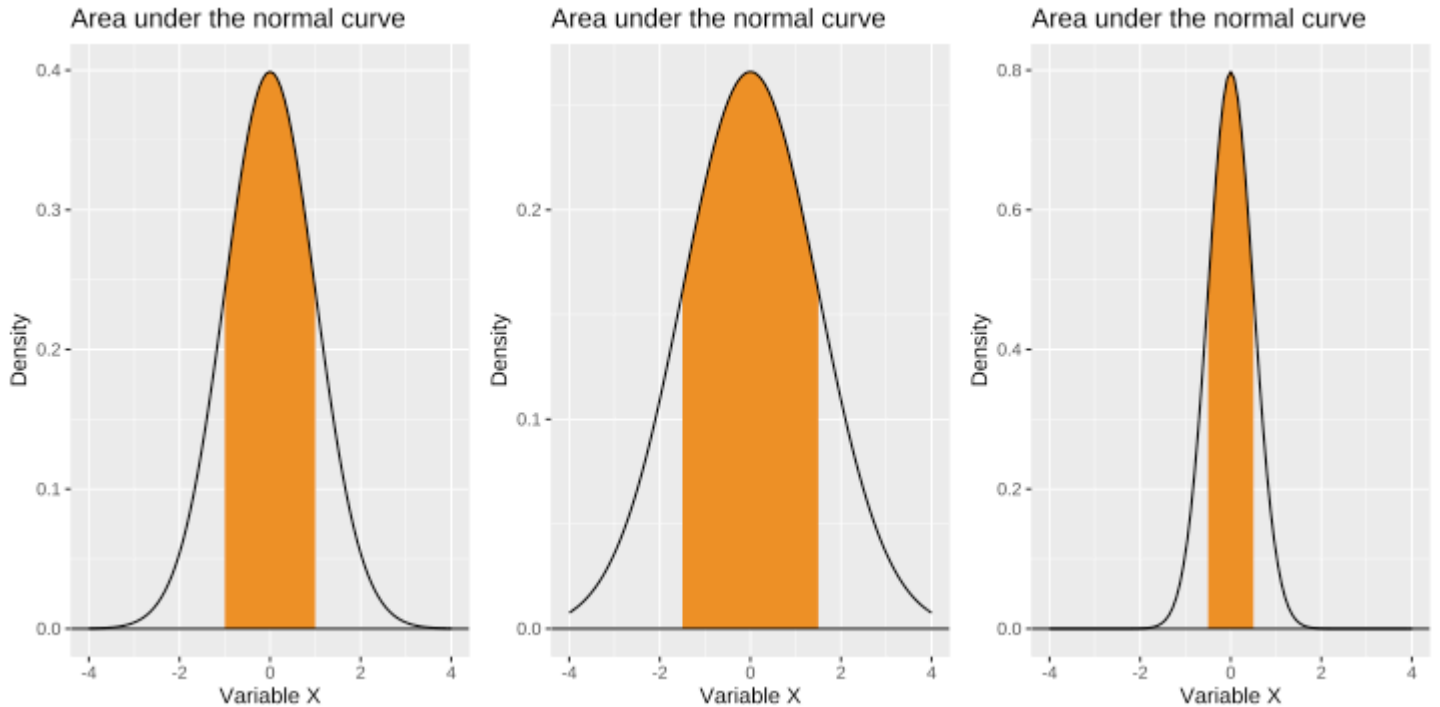


Characteristics of the normal distribution

- The mean and standard deviation are independent
- The distribution is unimodal and symmetrical
- For two normal distributions, the area under the curve between corresponding locations in standard deviation units is the same regardless of μ and σ

Family of Normal Distributions

All of these distributions are normal and have an equivalent area (proportion) that falls between a standard deviation below and above their respective means.



Characteristics of the normal distribution

- About 68.3% of the data will be within one standard deviation of the mean.
- About 95% of the data will be within two standard deviations of the mean.
- About 99.7% of the data will be within three standard deviations of the mean.

In other words, nearly **all** of the data will fall within 3 standard deviations of the mean in a normal distribution.

Standard normal distribution

A normal distribution with μ=0 and σ=1 is called **standard normal**. It is one specific distribution that comes from the larger family of normal distributions.

Variables with quite different means and standard deviations can be standardized so that they can be compared in the same metric (standard deviation units). This allows statements such as "relative to the mean, I am more conscientious (e.g., $z = 2$) than I am extraverted (e.g., $z = 1$)."

All continuous distributions can be standardized, but if they are not normal to begin with, standardization will not make them so. *Standardization does not alter distribution shape.*

Standardized scores (z-scores)

> Distance from the mean in standard deviation units

$$z = \frac{x_i - \bar{x}}{s}$$

Properties of z-scores:

- $\mu_z = 0$
- $\sigma_z = 1$
- Compare across scales and units of measures
- More easily identify extreme data

Using z-scores

```
## # A tibble: 6 x 2
##   name            height
##   <chr>            <int>
## 1 Luke Skywalker     172
## 2 C-3PO              167
## 3 R2-D2               96
## 4 Darth Vader        202
## 5 Leia Organa        150
## 6 Owen Lars          178
```

```
starwars %>%
  select(1:2) %>%
  mutate_at(2, ~round(x = scale(.
  head(.) %>%
  print(.)
```

```
## # A tibble: 6 x 2
##   name            height[,1]
##   <chr>                <dbl>
## 1 Luke Skywalker       -0.07
## 2 C-3PO                -0.21
## 3 R2-D2                -2.25
## 4 Darth Vader           0.79
## 5 Leia Organa          -0.7
## 6 Owen Lars             0.1
```

Using z-scores

Given any score, we can calculate the probability of getting a value greater than that z-score. *(Or less than that z-score.)*

You can look up tables that give you the probability value that corresponds to any given z -score. Or, you can use R code.

Luke Skywalker's height is z = -.07

```
pnorm(q = -.07, mean = 0, sd = 1)
```

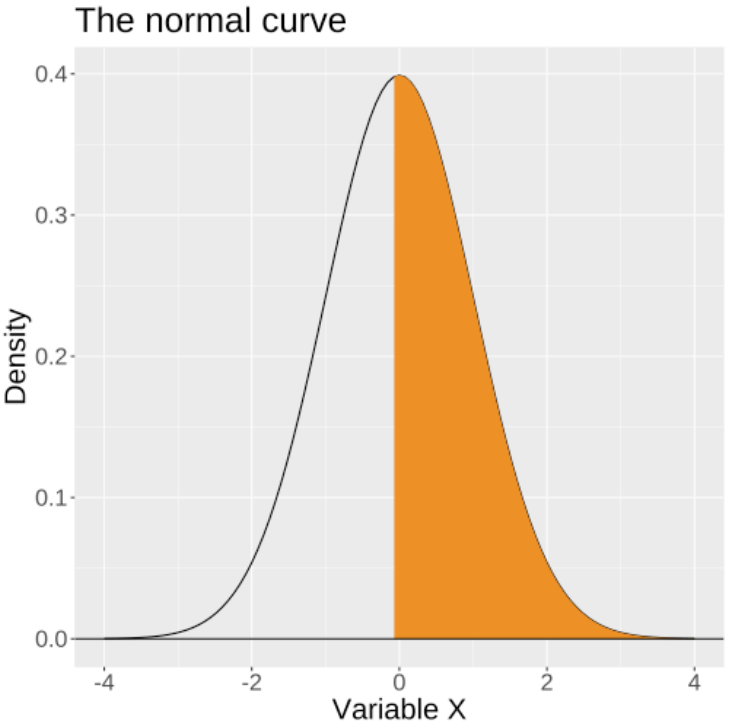
```
## [1] 0.4720968
```

Using z-scores

The probability of getting a z-score of -.07 or greater?

```
1-pnorm(q = -.07, mean = 0, sd =
```

```
## [1] 0.5279032
```



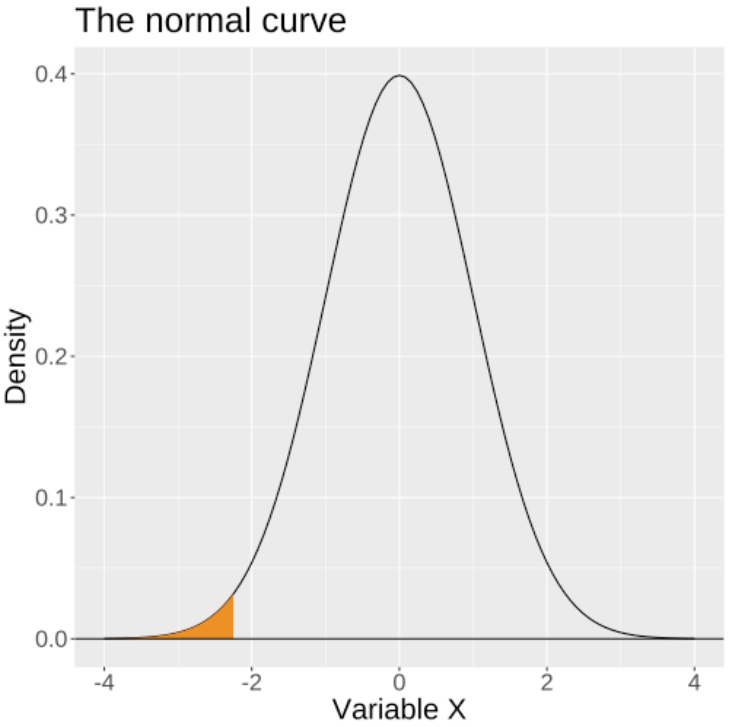
Using z-scores

What about R2D2? (z-score of -2.25)

- Probability of getting a z-score of -2.25 or something even smaller

```
pnorm(q = -2.25, mean = 0, sd = 1
```

```
## [1] 0.01222447
```

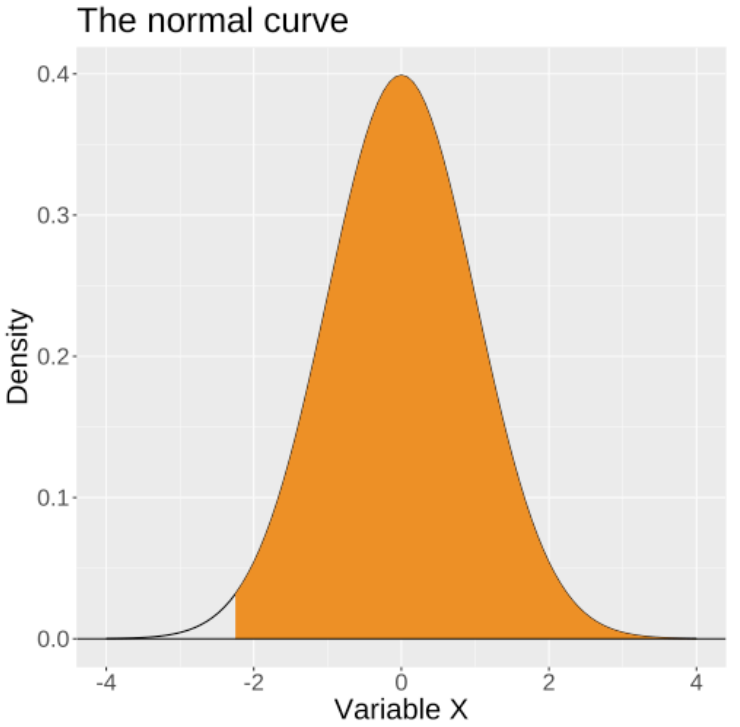


Using z-scores

Probability of getting a z-score of -2.25 or something larger

```
1-pnorm(q = -2.25, mean = 0, sd =
```

```
## [1] 0.9877755
```



Some z-scores of note

- $z = 1.64$; most extreme 5% of the standard normal distribution (the very far tail)
- $z = 1.96$; most extreme 2.5% of the standard normal distribution (used when splitting the difference of most positive and most negative extremes)

Sampling Distributions

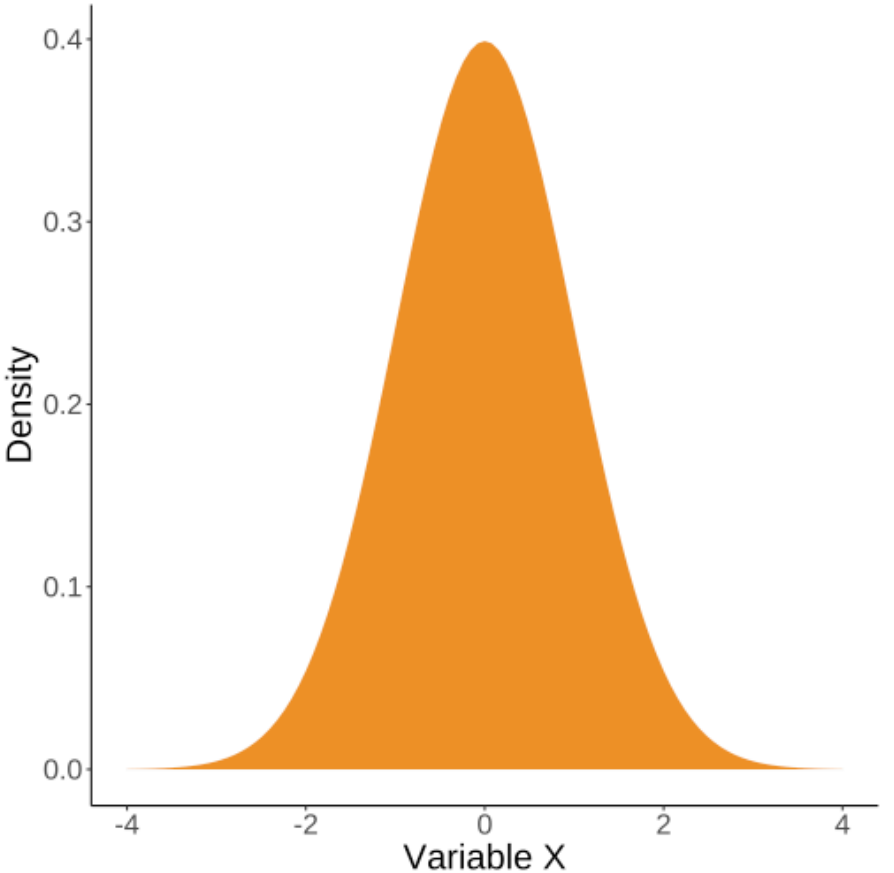
What's the point of inferential stats?

- Most of the time, we can't measure an entire population.
- We instead take random samples from the population, and we *estimate* statistics. We treat these as our best guess of the population parameter.
- We know that our statistics will vary from sample to sample.

	Population Distribution	**Sample Distribution**
Distribution consists of:	Individual observations x	Individual observations x
Central tendency	μ	$\bar{x}$
Dispersion	σ^2	s^2
	σ	s
Type	Parameter	Statistic
T vs. O	Theoretical	Observed

All sample statistics are wrong, but they become more useful as the sample size increases

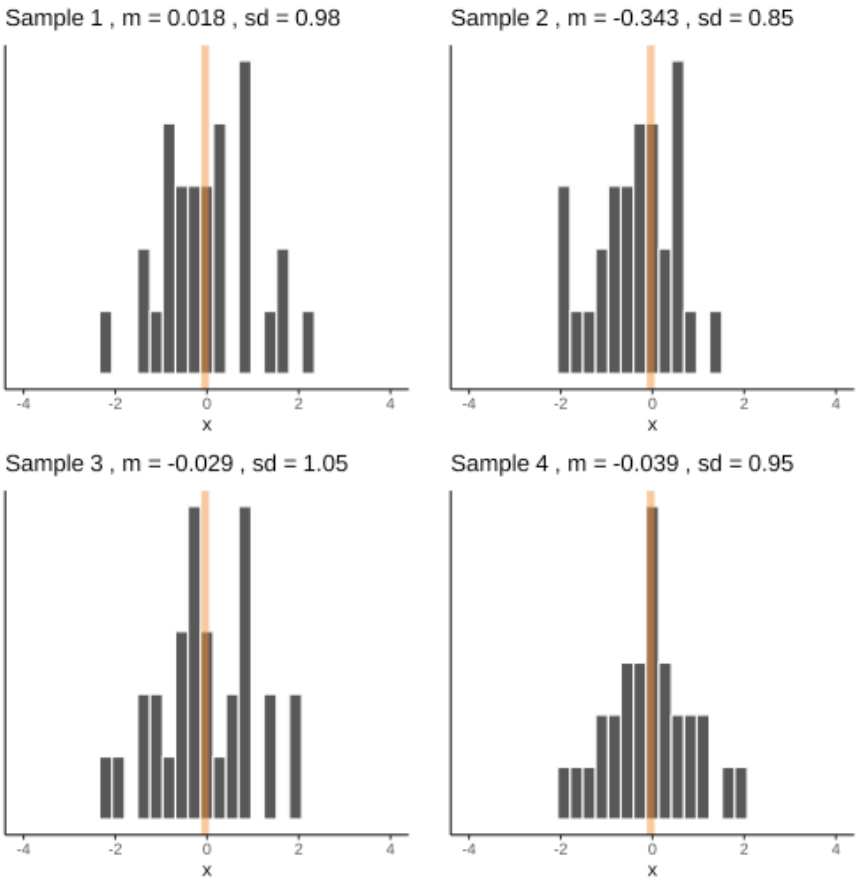
- The parameters of this population distribution are unknown
- We use the sample to inform us about the likely characteristics of the population



Samples from the population

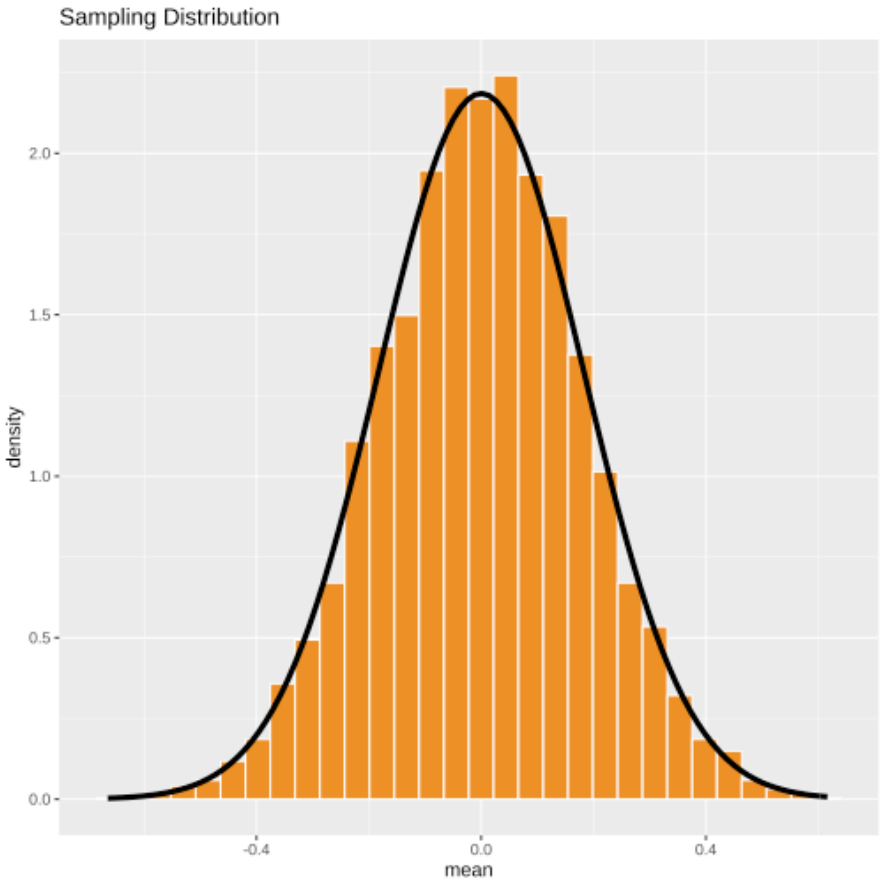
Each *sample distribution* will resemble the population. That resemblance will be better as sample size increases.

Statistics (e.g., mean) can be calculated for any sample.



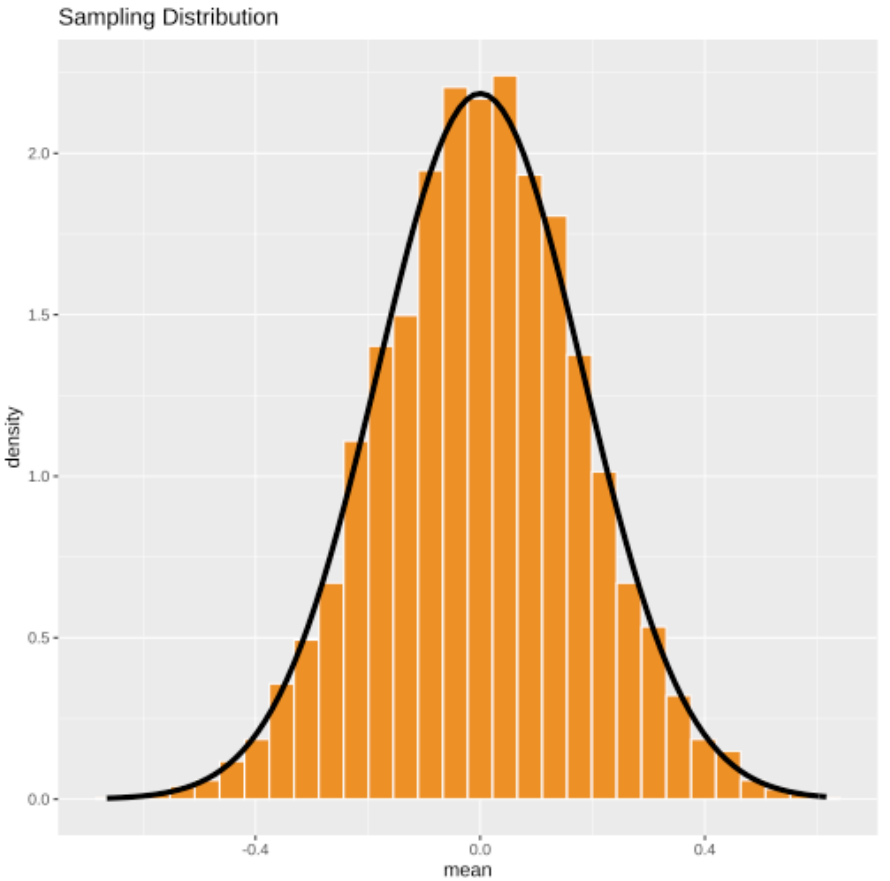
Sampling Distribution

- Say you repeated an experiment 100 times, each time using a new sample. Nothing else changes. Each sample has it's own mean (and other statistics).
- That's 100 means. You can have a distribution of means rather than of scores. That's a sampling distribution! **Distribution of statistics**



Sampling Distribution

- The mean of the sampling distribution converges on the population mean, μ
- The sampling distribution can also have it's own spread (variance/standard deviation). This tell us how typical or rare the sample statistic is likely to be. We call this the **standard error of the mean** (SEM).
- Note it could be any statistic (standard error of the median, standard error of the range etc.)



Notation

	Population Distribution	**Sample Distribution**	**Sampling Distribution**
Distribution consists of:	Individual observations x	Individual observations x	Statistics $\bar{x}, s, s^2$
Central tendency	μ	$\bar{x}$	μ_M
Dispersion	σ^2	s^2	σ_M^2
	σ	s	SEM σ_M
Type	Parameter	Statistic	Statistic of statistics
T vs. O	Theoretical	Observed	Theoretical

Sampling Distributions

- Distribution of values of a particular statistic ($\bar{x}$, s^2, s) across all possible samples of N observations
- One of the most important discoveries in statistics is that the sampling distributions of many statistics are approximately **normal** even when the sample (and population) distributions are not normal!
- Play around with this if you want to prove it to yourself
- Why does this matter? Because as we saw earlier, the normal distribution is awesome!

Scroll back

- Remember that whole exercise we did earlier in this lecture with the Starwars dataset?
    - We took a vector of heights
    - Turned them into z-scores
    - Asked *"how likely is it that we got this particular z-score or something more extreme?"*
- Now, we are going to do this exact sampe procedure, but this time, rather than working with individual scores, we're going to work with **means**. Is it likely or unlikely that we got a particular mean (or more extreme), if it comes from a particular sampling distribution of means?

This is what we're actually interested in!

Next time...

- Comparing means with NHST...putting it all together