Lecture 7: Towards statistical models

Descriptive statistics and the sampling distribution

Dr Milan Valášek

9 November 2020

UNIVERSITY OF SUSSEX

Overview

Measures of central tendency

- Mode
- Median
- (arithmetic) Mean

Measures of spread

- Range
- Interquartile range
- Variance and standard deviation

Going meta

- Sampling distribution
- Standard error

Describing things with maths

- Quantitative methodology deals with measurable things (variables)
- It explains and predicts the world around us by modelling relationships between variables
- These models are mathematical/statistical in nature and they are based on numeric descriptions of variables
- Variables differ in their range and distribution and *from population to population*
  - Air temperature on Earth ranges from about –90°C to about 60°C
  - Temperature produced by humans under laboratory conditions: –273°C - 5.5 trillion°C
  - Distribution of height is *normal*, distribution of wealth is *skewed*
- The term population does not only refer to people!
- The most basic ways in which we can describe variables and their distributions is in terms of central tendency and spread

Central tendency and spread

- Distribution of the values in a variable can be described in terms of
  - its "average" value, *i.e.*, where the *"most typical"*, or central value is located along the possible range of values
  - how much *variability* there is in the individual values of the variable in the sample or population , *i.e.*, how much the values are spread along the range of values
- There are various measures of both central tendency and spread, each with its pros and cons
- All of them are mathematical *abstractions* - they provide useful information but they're not all there is to things

Measures of central tendency

- Measures that tell us about the *"most typical"* value of a variable
- What does "most typical" mean though?
- Different measures of central tendency have different answers

Plot | R code

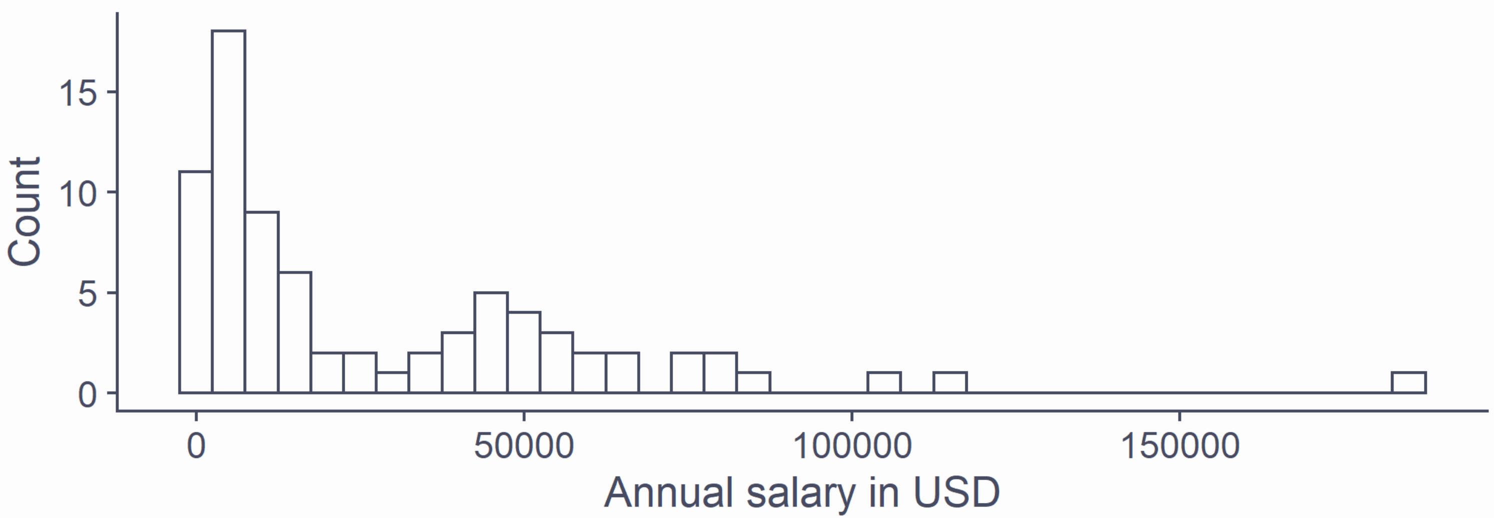


Fig 1 Average national annual salary on a sample of 78 countries

Measures of central tendency

- We'll talk about three of these measures
  - the mode
  - the median
  - the arithmetic mean
- They are all different kinds of *average*

Mode

- The most frequent value in the distribution
- A distribution can have one or more modes

Plot | R code

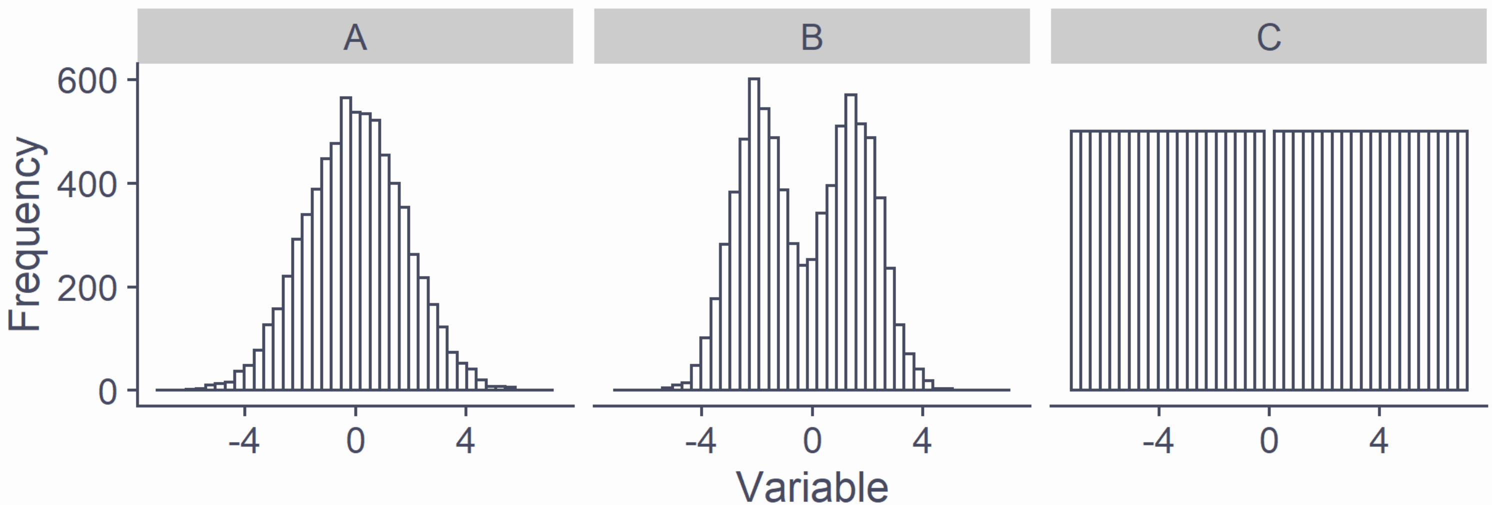


Fig 2 Examples of a (A) unimodal, (B) bimodal, and (C) multimodal distribution

Median

- To find the median, first sort data
- Then find the mid-point (average of two mid-points if the number of observations is even)

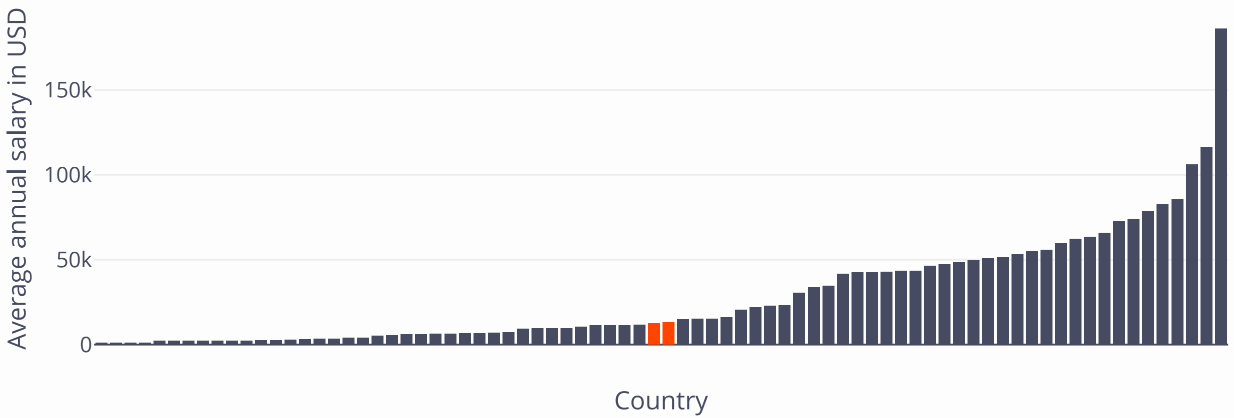


Fig 3 Average national annual salary per country sorted from lowest to highest

```
median(salary$yearly)
```

```
## [1] 12855
```

Mean

- What most people mean by *average*
    - population mean μ
    - sample mean $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- If there are *N* observations of variable *x* in our sample,

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + x_3 + \cdots + x_N$$

```
mean(salary$yearly)
```

```
## [1] 28685.77
```

Median

- Not influenced by extreme values in sample (50% of sample is larger and 50% smaller, no matter what)
- Does not have a proper algebraic formula
- Medians of different samples from the same population can be relatively different from each other

Mean

- Has a formula which allows us to do all sorts of maths (and stats) with it
- Means of different samples from the same population are relatively similar to each other
- Sensitive to extreme values
- Basis for some measures of spread

Mean Vs Median

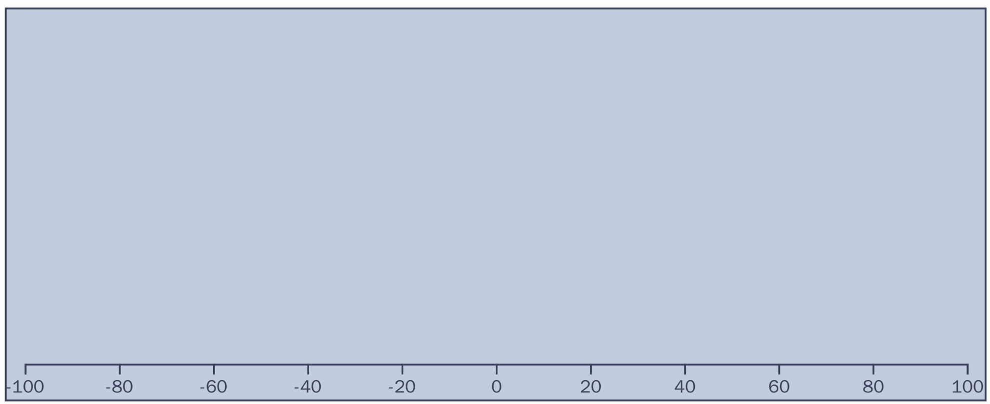
Location Spread

Reset

Median

Mean

+5



Variable types and central tendency

Mode

- Mainly for discrete variables
- Doesn't make much sense for truly continuous variables

Median

- For variables that can be measured on *at least the ordinal level*

Mean

- For variables that can be measured on *at least the interval level*

Measures of spread

- Mode, median, and mean tell us about the central point of a variable
- They don't tell us how spread the data are around this point, *e.g.*, how much variability there is in the variable

Plot | R code

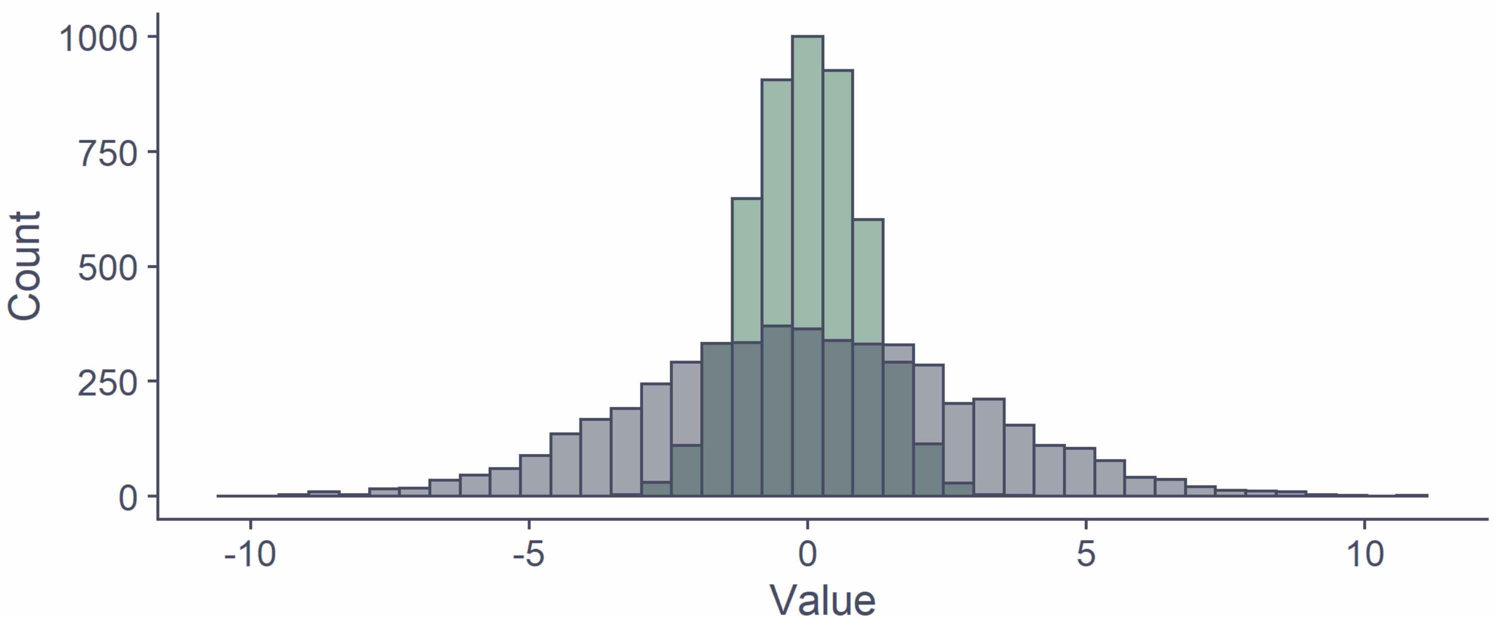


Fig 3 Histogram of two distributions with equal means but different spread. *N*=5,000 in each case.

Measures of spread

- Measures of spread (dispersion) tell us about the variability in the data
- We will look at the following:
  - Range
  - Inter-quartile range
  - Deviation
  - Variance
  - Standard deviation

Range and Inter-quartile range

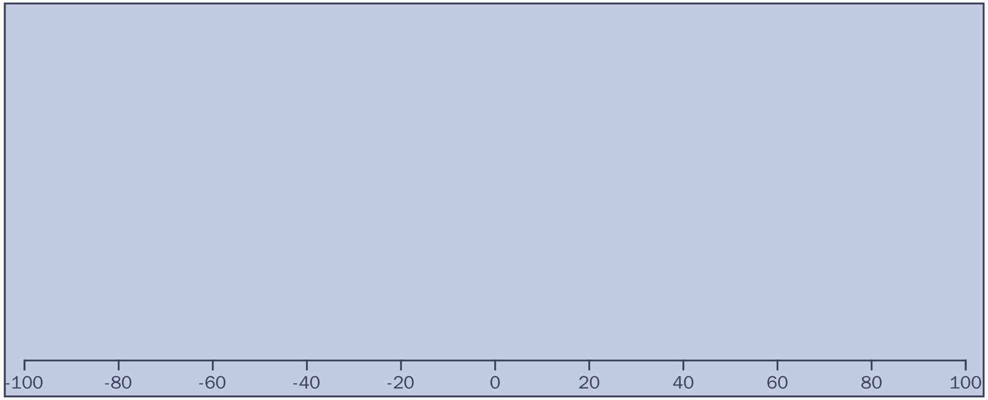
Location Spread

Reset

Median

Mean

+5



Range

- Distance between smallest and largest value in sample
- *Drawback:* Extremely sensitive to outliers

```
max(salary$yearly) - min(salary$yearly)
```

```
## [1] 185560
```

IQR

- Inter-quartile range - distance between 1st and 3rd quartile
- *Drawback:* Ignores half of the data

```
IQR(salary$yearly)
```

```
## [1] 41820
```

Deviation

- Distance from every single value in the data from some convenient point
- Mean is a convenient point
- $x_i - \bar{x}$, where x_i is every single data point
- There are as many deviations as data points
- To get a single measure of spread, how about we add up the deviations?
- *Problem:* More data points = more points to add up
- *BIG problem:* They always add up to zero

Deviation and variance

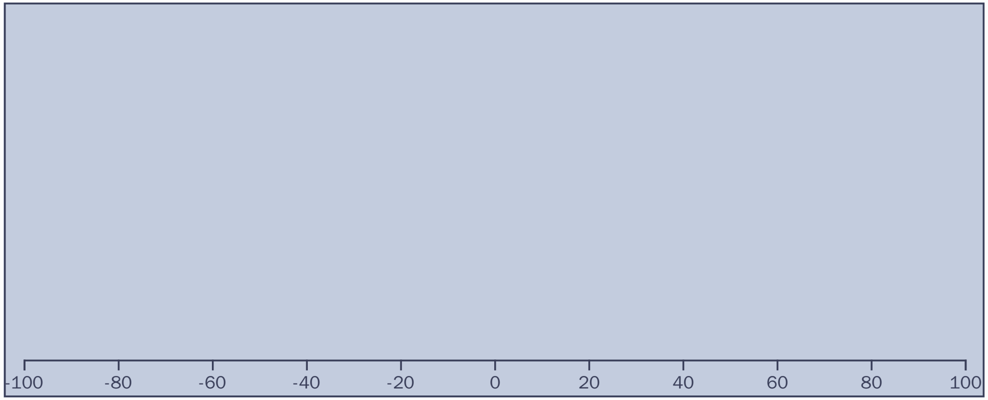
Location Spread

Reset

Median

Mean

+5



Variance

- We get around the *BIG problem* (deviations adding up to 0) by taking the square of the deviations
    - The sum of these is called the *Sum of Squares*
- We can get around the *problem* by dividing the sum of squares by *N*
    - This is the variance

Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Sample variance: $s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$

- reason for the $N-1$ is quite technical (see Bessel's correction)

```
var(salary$yearly)
```

```
## [1] 1095256212
```

Standard deviation

- Variance is a good measure of dispersion and is widely used
- One minor inconvenience is that it's measured in *squared units*
    - if salary is measured in years, s^2_{salary} is expressed in USD^2, whatever those are
- Taking the square root of variance gives us a measure of spread in the original units
- This is the standard deviation
    - σ for population
    - *s* (or *SD*) for sample

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}}$$

```r
sd(salary$yearly)
```

```
## [1] 33094.66
```

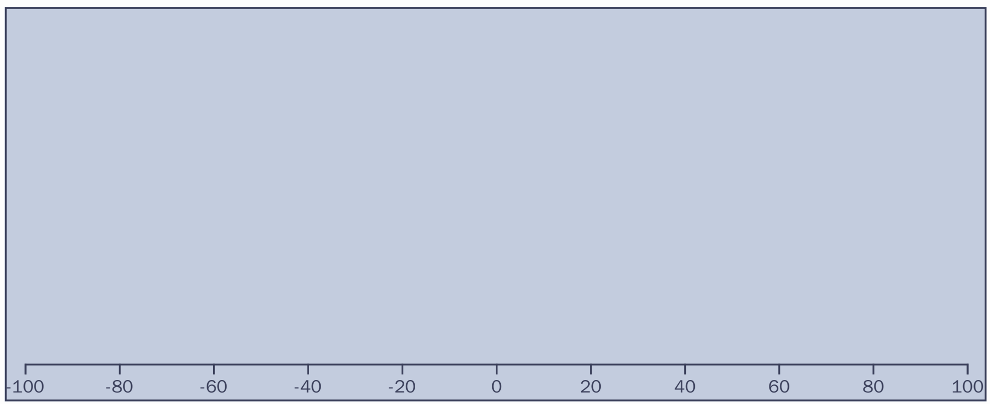
- If we measured salary in 1000s of USD, $s_{\text{salary1000}}$ would be proportional to s_{salary} (33.09 as opposed to 33094.66)

Standard deviation

Location Spread Reset

Median

Mean +5



From sample to population

- We want to make claims about the world
- *We don't care about samples, we care about populations*
- However, we cannot measure the entire population so we have to make do with samples
- So we end up making claims about the world based on what we know from the sample
- We *cannot be sure* that our sample accurately represents the population
- Because of that, there's always uncertainty associated with any empirical claims we make

From sample to population

A full set of Scrabble tiles contains 100 tiles with a mean tile value of 1.87 points and a *SD* of 1.83.

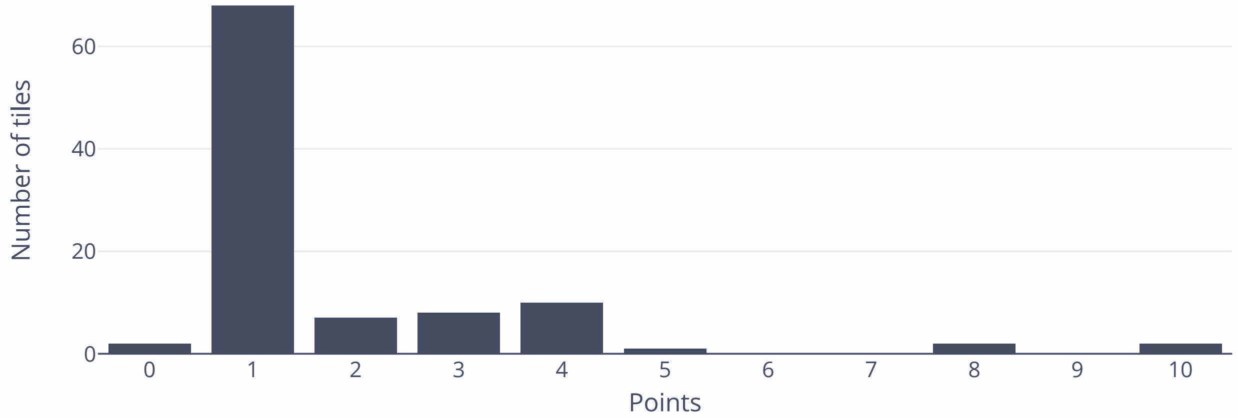


Fig 4 Distribution of Scrabble tiles by point value.

You draw - sample - 7 tiles for your rack

- Sometimes you pick only vowels, sometimes you get only the Zs, Qs, Ks, or Ws due to *statistical fluctuation* in sampling
- Most often, you pick a mix of low-score and high-score tiles

Draw

►

Reset

Plot

Sampling distribution

Sampling distribution is the distribution of a statistic (*e.g.*, the mean) based on all possible samples of a given size taken from the same population

- In the Scrabble example it's the distribution of all possible means of 7-tile draws.
- *Sampling distribution is* NOT *the distribution of the sample!*
- The centre (mean) of the sampling distribution is equal to the population value of the calculated statistic
  - The mean of the sampling distribution of the mean *is equal to the population mean*
  - The mean of the sampling distribution of variance *is equal to population variance*
- The standard deviation of the sampling distribution is called the standard error (*SE*)
  - Very important concept!
  - Allows us to quantify the uncertainty about our estimates

Standard error

- *SE* is the standard deviation of the sampling distribution
- Quantifies the uncertainty about how similar the sample statistic (*e.g.*, the sample mean, $\bar{x}$) is likely to be to the population parameter (*e.g.*, population mean, μ)

$$SE = \frac{\sigma}{\sqrt{N}}$$

- Related to sample size and variability in population
  - If mean annual salary doesn't change much from country to country, *SE* will be relatively small
  - If our sample is large, *SE* will be relatively small and *vice versa*
- The concepts of the sampling distribution and standard error will be of crucial importance later, when we are talking about testing hypotheses and statistical modelling

Recap

- We can describe distributions ("shapes of variables") using maths
- Central tendency refers to the mid-point of a variable
  - Mode
  - Median
  - Mean
- Spread refers to the amount variability in the variable
  - Range
  - IQR
  - Variance
  - Standard deviation
- Each measure has its properties and is useful in different situations

Recap

- We don't care about samples, we care about populations
  - But we have to rely on on samples because we don't have access to populations
- Different samples have different properties (*e.g.*, means) even though they are sampled from the same population
- The sampling distribution is the distribution of a given statistic from all possible samples of the same size drawn from the same population
- The standard deviation of the sampling distribution is the standard error
  - *SE* quantifies the uncertainty about how similar the sample statistic is to the population parameter
  - The larger the sample, the smaller the *SE*
  - More variable populations lead to larger *SE*s


That's all Folks!