Unit 4: Sampling

3. Sampling distribution

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

From section "Health care utilization and days of work or school missed (sick days)"
"In questionable cases, classification as ARI-related or not was done by the senior author (B.B.), guided by relevant data but blinded to allocation."

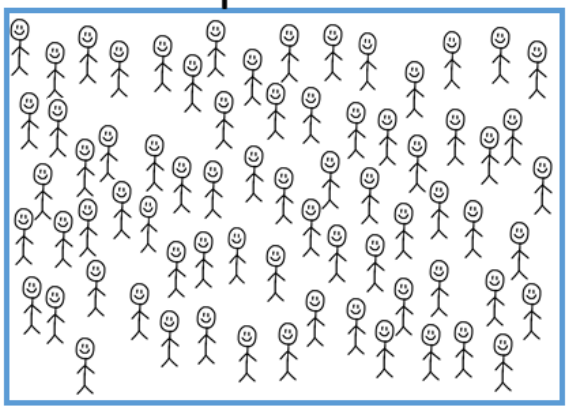
This is a discussion on how the response variable (improving respirator disease) is measured. In general, when we talk about blind/double-blind, we specifically mean blinding the participants to the treatment.

The treatment is here assigning the participants to three study groups: (1) exercise, (2) meditation, (3) control group. Both subjects and researchers knew what treatment each participant was assigned to. Therefore, no blinding is implemented.

- EA04 is the first part of your final project (EA04-07)
- Form a team of 4 students, and submit your group information on Piazza
- Each of you will have to be the team captain for once in your final project
- Start thinking about what data you would like to work on for the final project, and what research questions can you answer with your data?
- More will be discussed on Friday

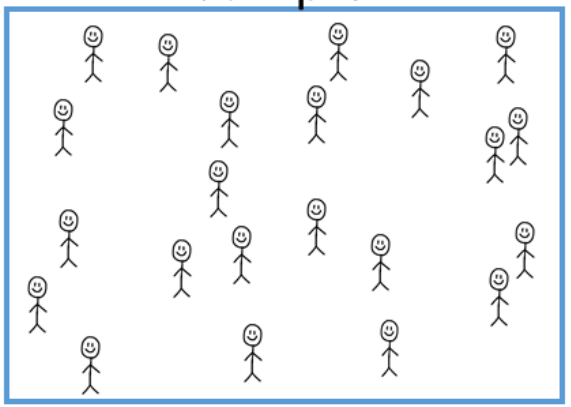
Outline

Population



Take Sample

Sample



Calculate Sample Statistic

$\hat{p}$: **sample proportion** – what proportion of people in the **sample** support the president?

Inference

Our Goal: Estimate the Population Parameter

p: **population proportion** – what proportion of people in the **population** support the president?

Outline

I recently received a gift of 100,000 M&M's.

Since my favorite color is blue (Guess why?), I would like to know what proportion of these M&M's are blue?

One way to answer this question would be to perform an exhaustive count. However, this would be a long and tedious process. Instead, let's do some sampling activity!

1. What is the population of study?
2. What is the question that we are interested in here?
3. What is the population parameter?
4. What symbol will you use for the population parameter?

Instead of performing an exhaustive count, let's remove a small portion of M&M's from the gift (say 50 M&M's)

Because of the remote setup of this course, let's do a virtual sampling in R.

1. What is the sample here?
2. What is the sample statistics?
3. What symbol will you use for the sample statistics?

Now think...

Why couldn't we just use one sample?

Now each of you, please use the provided R code in the template to sample 50 M&M, and log in the shared excel sheet the proportion of blue M&M's in your sample.

Did you all have the same proportion in your sample? Why?

Let's plot the histogram of the numbers you put in the excel sheet...

- Where is the center of the histogram?
- What's the shape of the histogram?

This is getting boring. Can we automate the process?

Code

```
sample_props50 <- mm %>%
  rep_sample_n(size = 50, reps = 24, replace = TRUE)
```

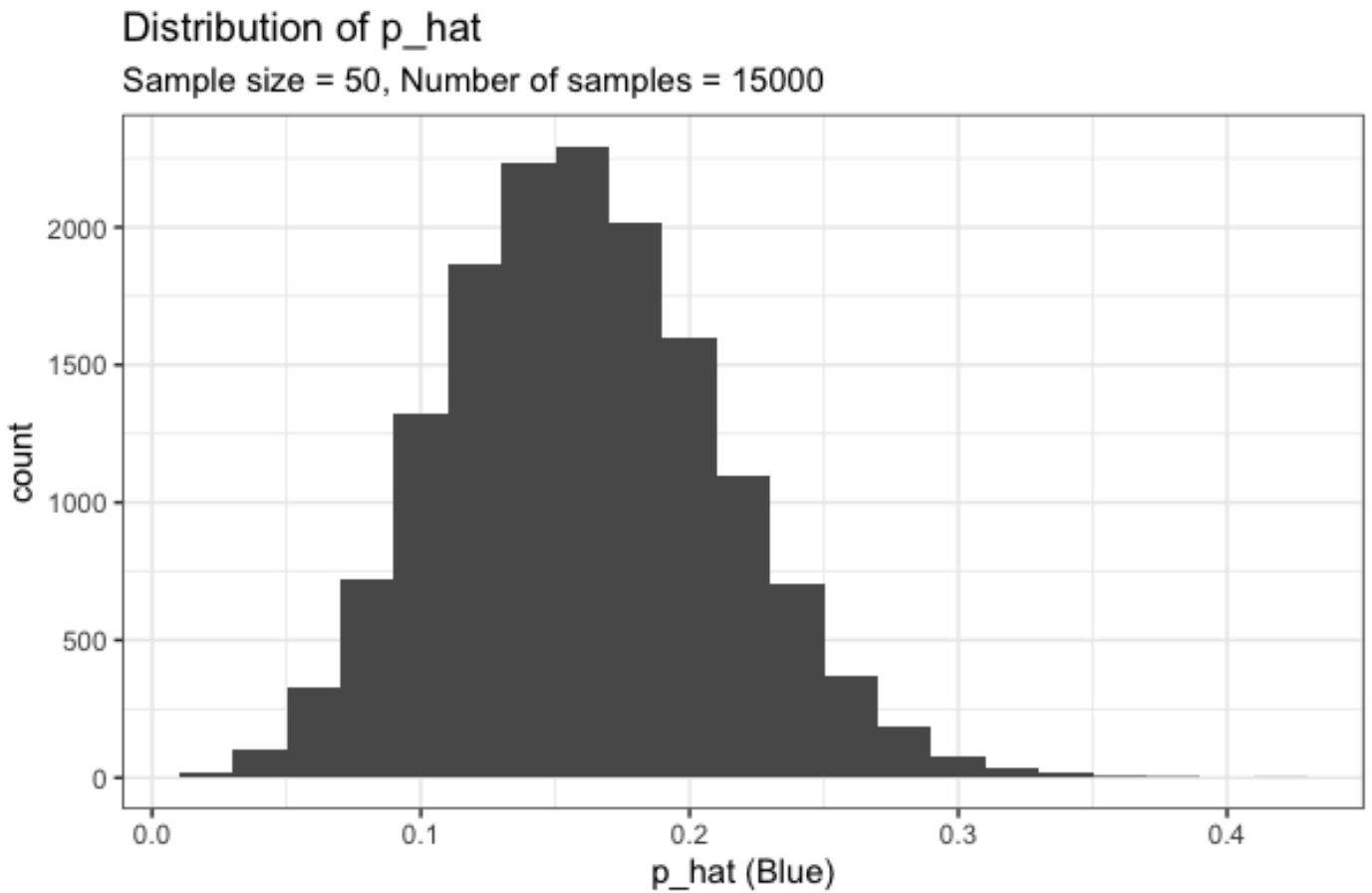
The function rep_sample_n took a random sample of size n (50) from the population of all M&M's and repeat this sampling procedure rep (24) times.

Note that we specify that replace = TRUE since sampling distributions are constructed by sampling with replacement.

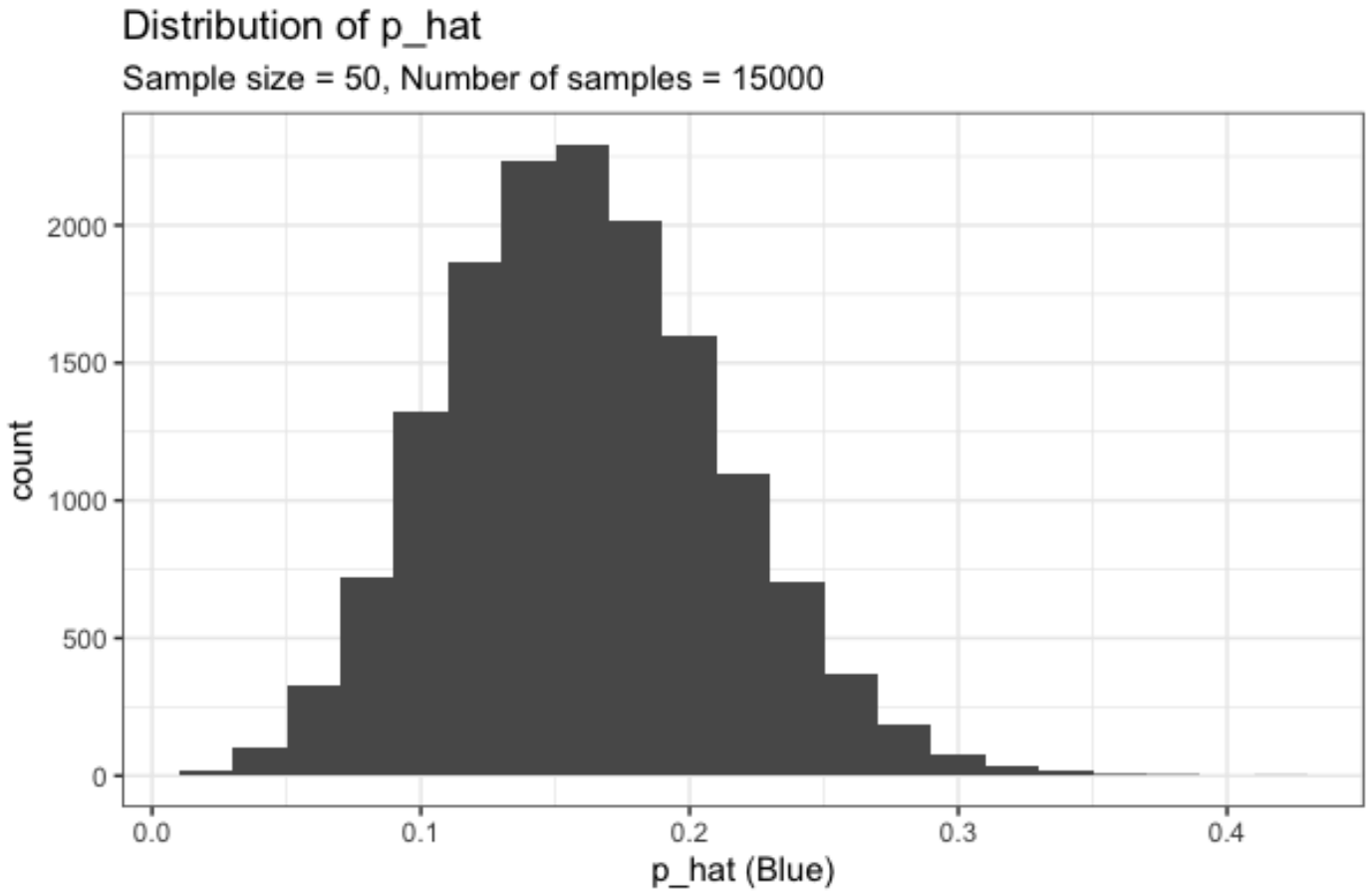
1. What does the x-axis represent in this histogram?
2. What does the y-axis represent in this histogram?
3. How does this compare to the first histogram we plotted?
4. Why do the two histograms look different?

Code

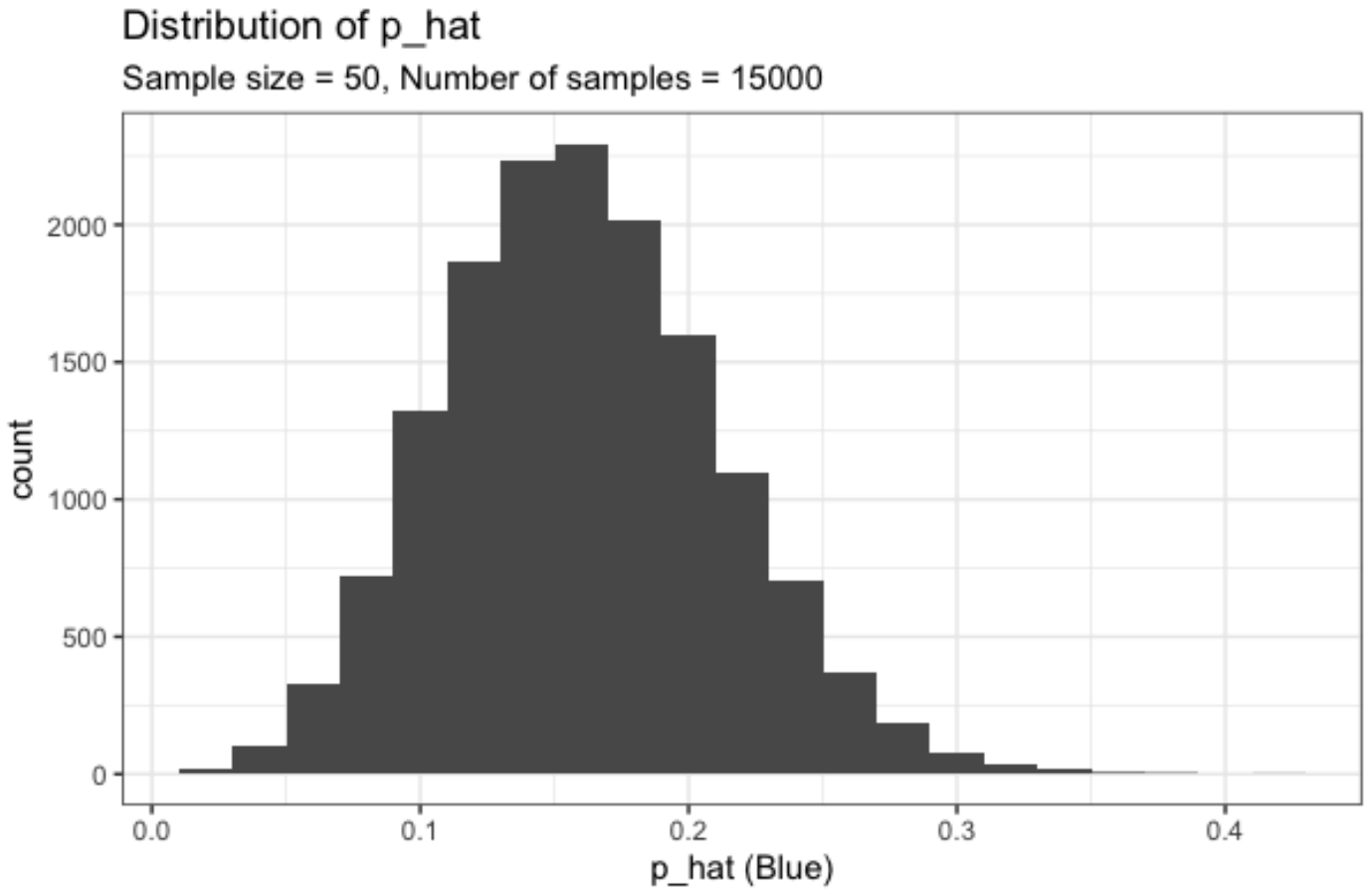
```
sample_props50 <- mm %>%
  rep_sample_n(size = 50, reps = 15000, replace = TRUE)
```



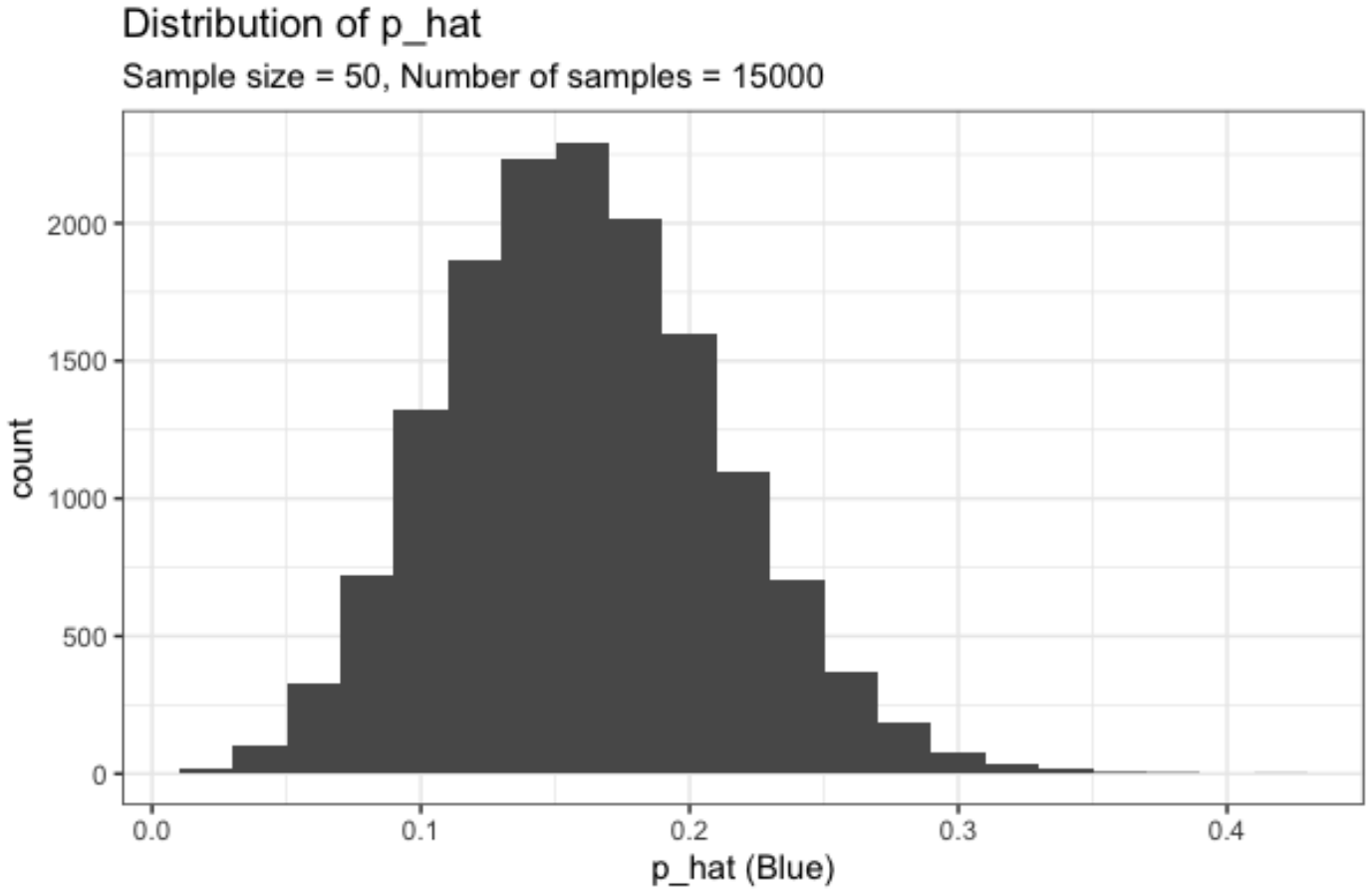
Would you say that sampling 50 M&M's where 15% of them were blue is likely or not? What about sampling 50 M&M's where 25% of them were blue?



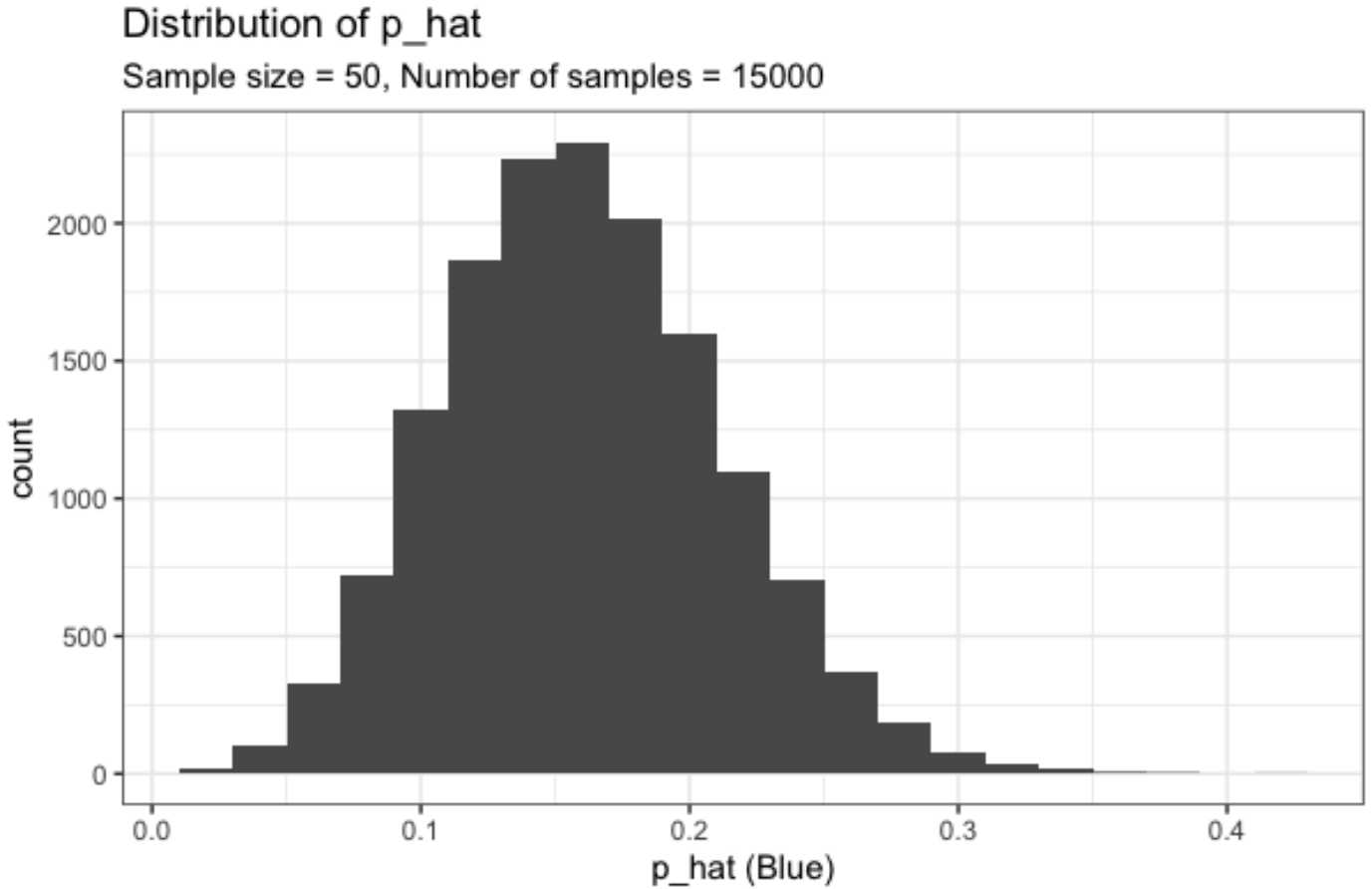
Based on this histogram, what's your estimate for the proportion of blue M&M's in the entire 100,000 M&M's gift?



How to interpret the spread of the histogram?



How to interpret the spread of the histogram?



The spread of the distribution indicates how much variability is incurred by sampling only 50 M&M's at a time from the population

You have three choices of cups to extract a sample of balls with: cups of size 25, 50, and 100.

If your goal is still to estimate the proportion of the bowl's balls that are red, which shovel would you choose?

In the provided code chunk, set the parameters sample_size to be 50, 100 and 1000, respectively, while keeping the number of repeated/replicated samples at 15000.

Comparing the histograms you got, what change have you noticed?

1. Take a random sample taken with replacement
2. Calculate the sample statistic - a statistic such as mean, median, proportion, slope, etc
3. Repeat steps (1) and (2) many times to create a sampling distribution - a distribution of sample statistics.
4. Use the center of the sampling distribution to estimate the true population parameter.

A **point estimate** is a single value computed from the sample data to serve as the "best guess", or estimate, for the population parameter.

Outline

1. Every sample you take is different!
2. So each sample will have a different sample statistic
3. The **sampling distribution** is the distribution of values of the sample statistic that you would get from all possible samples of a given size.

An important idea:

How the value of a sample statistic would vary from sample to sample, if random samples were randomly selected over and over from a population