Week 4: Statistical theory

3. Sampling distribution

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

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

Outline

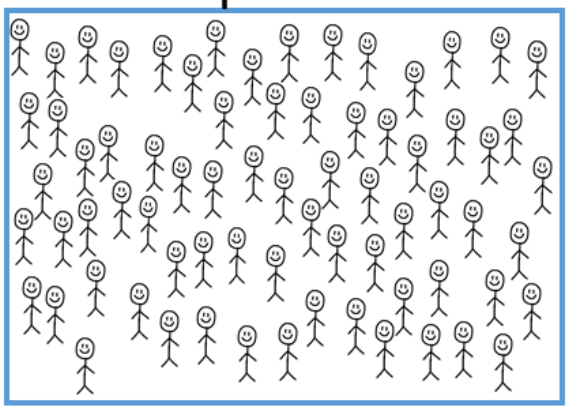
1. Today: sampling distribution

2. A sampling activity

3. Main take-away

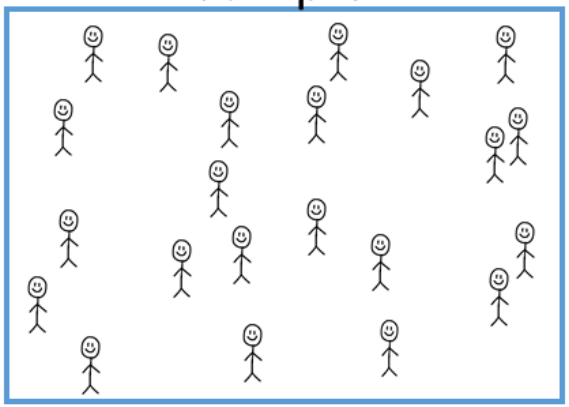
The important picture

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?

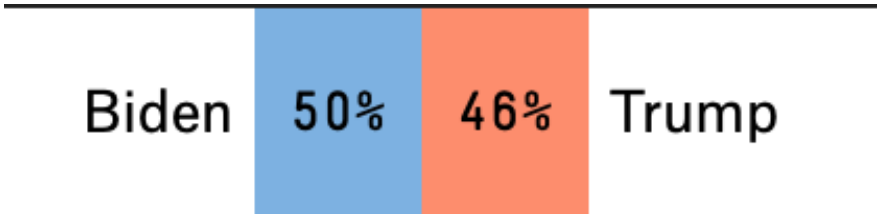
A **parameter** is a number that describes some aspect of a population.
A **statistic** is a number that is computed from data in a sample.

We usually have a sample statistic and want to use it to make inferences about the population parameter

Common population parameters of interest and their corresponding sample statistic:

Quantity	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	p	$\hat{p}$

Before the 2020 presidential election, 1000 registered voters were asked who they plan to vote for in the 2020 presidential election.



What proportion of voters planned to vote for Biden?

$$\hat{p} = 0.5, \qquad p = ?$$

We use the statistic from a sample as a **point estimate** for a population parameter.

Point estimates will not match population parameters exactly, but they are our best guess, given the data.

Actually, several polls were conducted over this time frame (10/20/20 – 10/26/20)

DATES	POLLSTER	SAMPLE	RESULT
OCT 22-26, 2020	C+ Rasmussen Reports/Pulse Opinion Research	1,500 LV	Biden 49% 47% Trump
OCT 22-26, 2020	A/B IBD/TIPP	970 LV	Biden 50% 46% Trump
OCT 22-26, 2020	A/B IBD/TIPP	970 LV	Biden 51% Less ⊗ Trump 46% Jorgensen 1% Hawkins 1%
OCT 13-26, 2020	B/C USC Dornsife	5,293 LV	Biden 53% 43% Trump
OCT 13-26, 2020	B/C USC Dornsife	5,293 LV	Biden 54% 42% Trump
OCT 13-26, 2020	B/C USC Dornsife	5,197 RV	Biden 52% 42% Trump

Sample statistics vary from sample to sample. (they will not match the parameter exactly)

KEY QUESTION

For a given sample statistic, what are plausible values for the population parameter? How much uncertainty surrounds the sample statistic?

KEY ANSWER

It depends on how much the statistic varies from sample to sample!

Outline

1. Today: sampling distribution

2. A sampling activity

3. Main take-away

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!

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 population parameter?
2. What symbol will you use for the population parameter?
3. What is the sample statistics?
4. What symbol will you use for the sample statistics?

A sampling distribution is the distribution of sample statistics computed for different samples of the same size from the same population.

A sampling distribution shows us how the sample statistic varies from sample to sample

Poll question

In the M&M's sampling distribution, what does each dot represent?

1. One M&M
2. One sample statistic/point estimate

This is getting boring. Can we automate the process?

Code

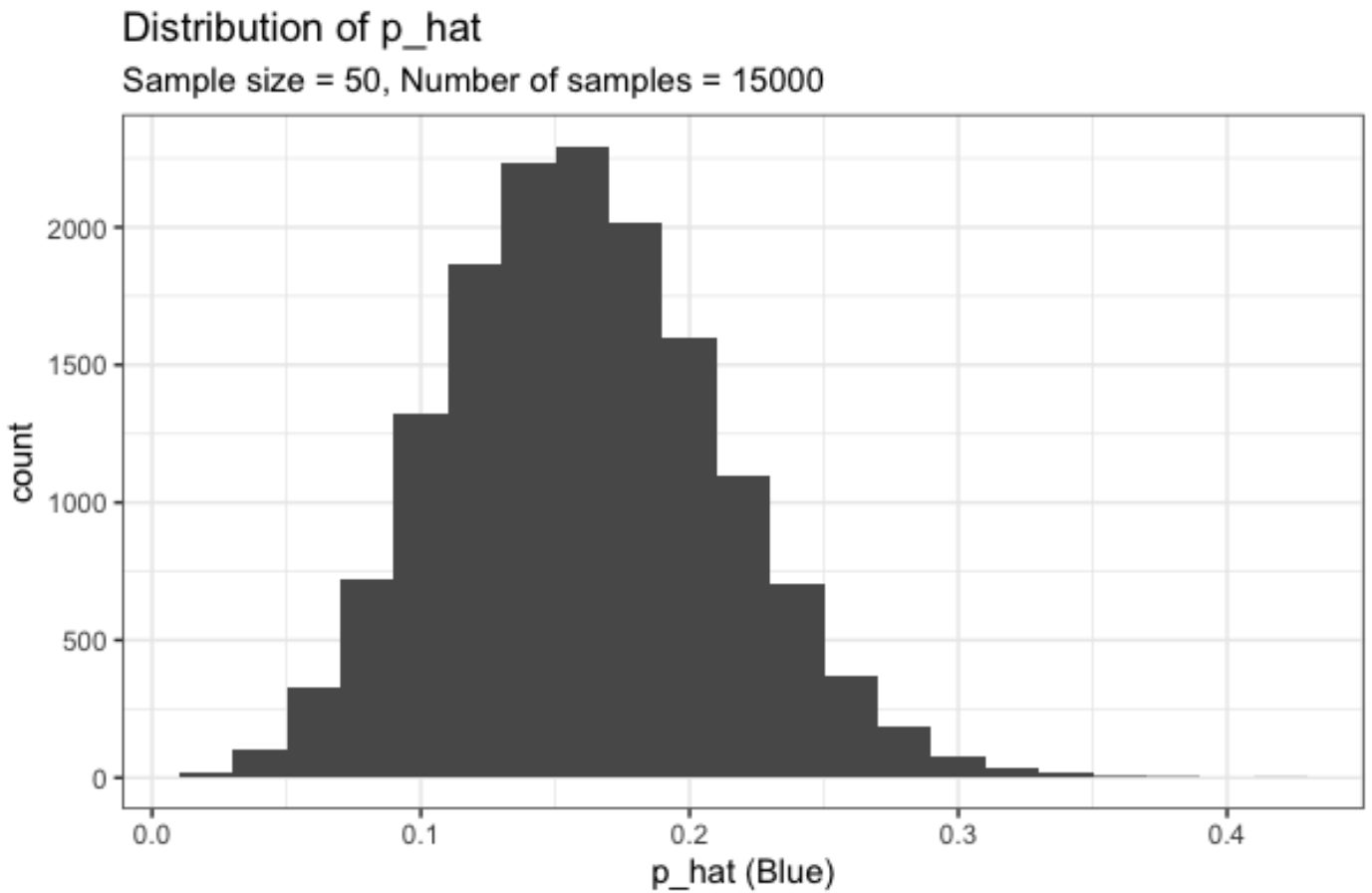
```
sample_props50 <- mm %>%
  rep_sample_n(size = 50, reps = 24, replace = TRUE)
```

The function rep_sample_n took a random sample of size (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.

Code

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



Center

If samples are randomly selected, the sampling distribution will be centered around the population parameter.

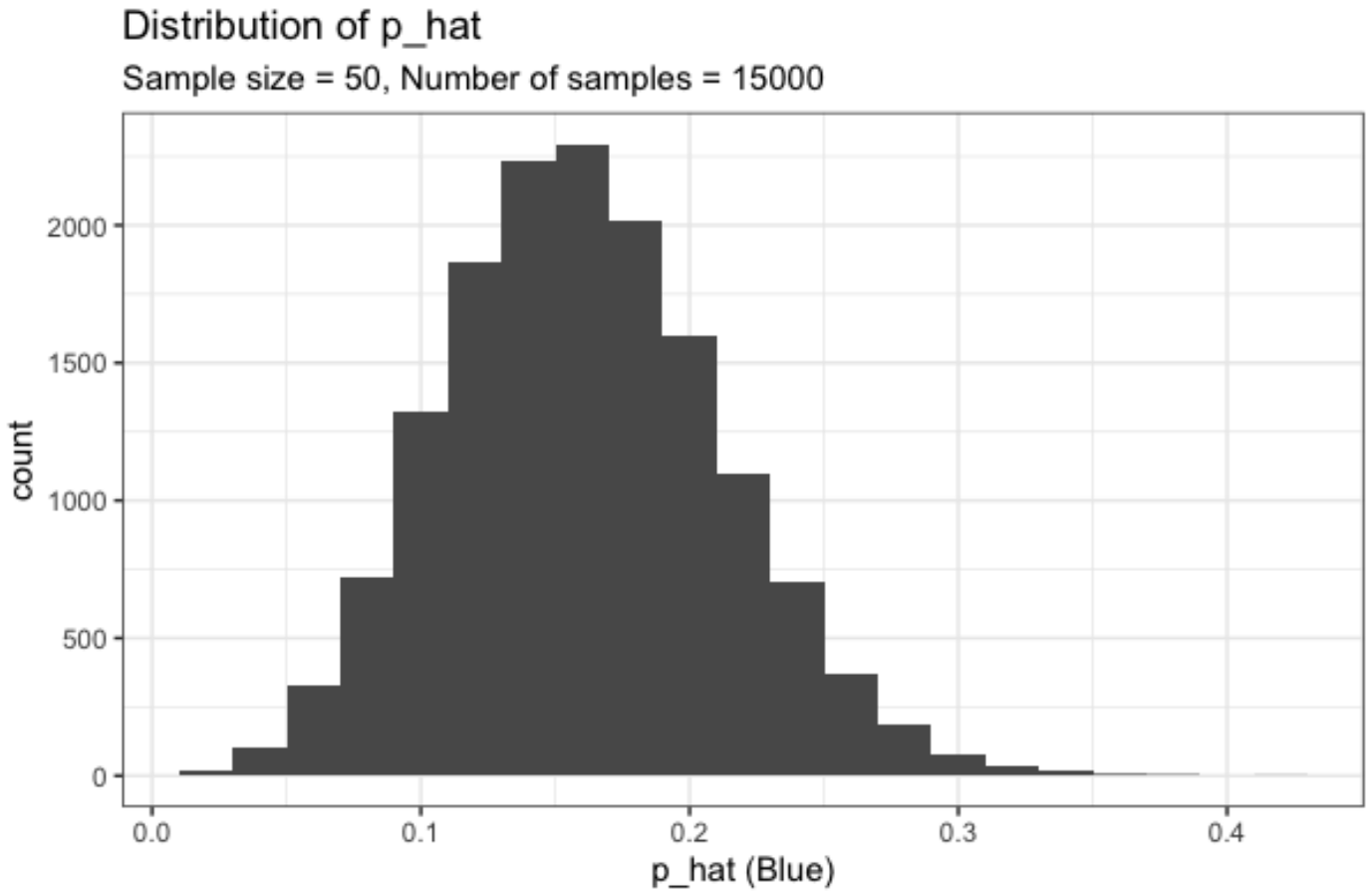
Spread

The spread of the distribution measures how much the statistic varies from sample to sample. Also known as **Standard Error**

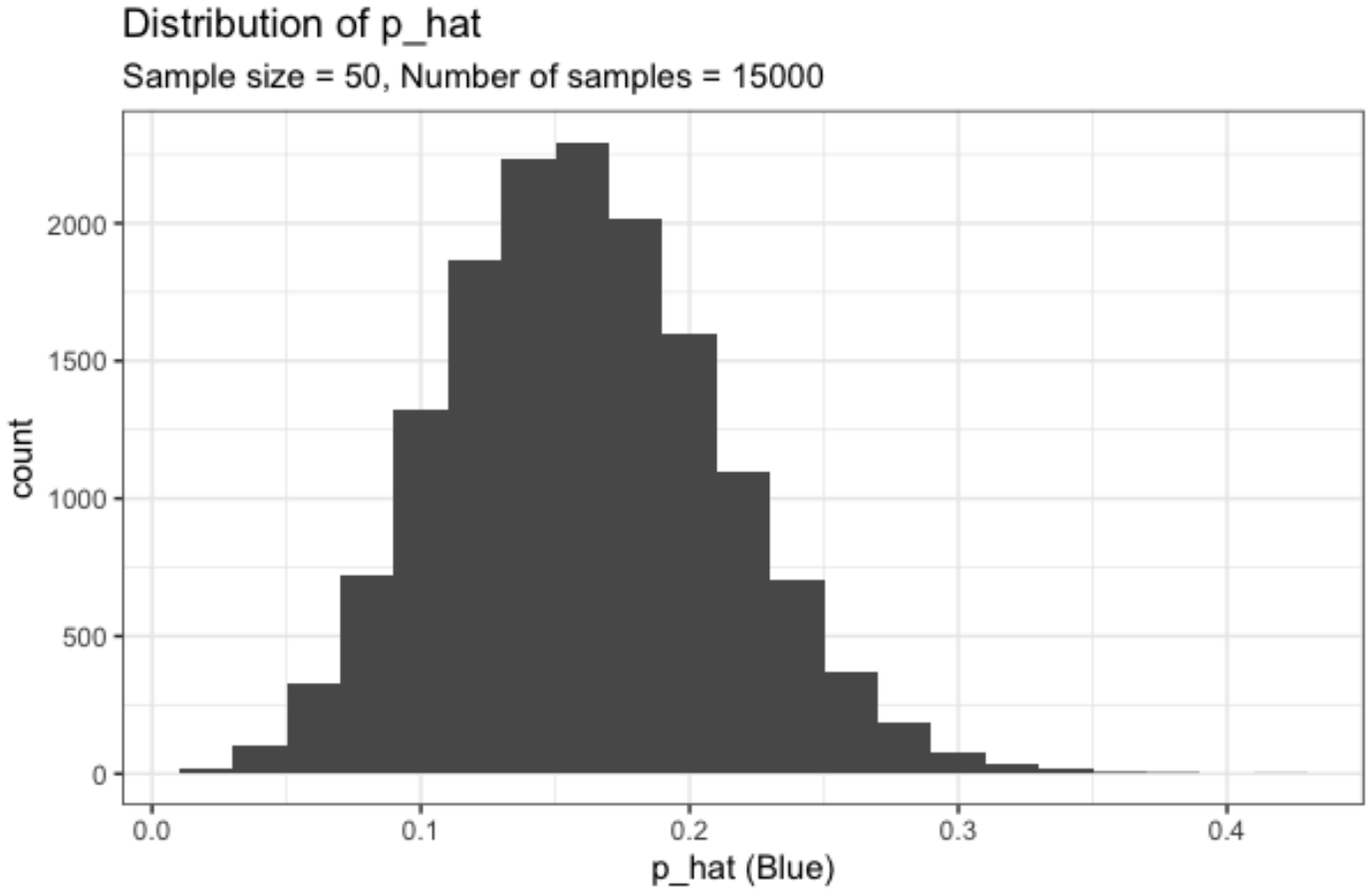
Shape

For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.

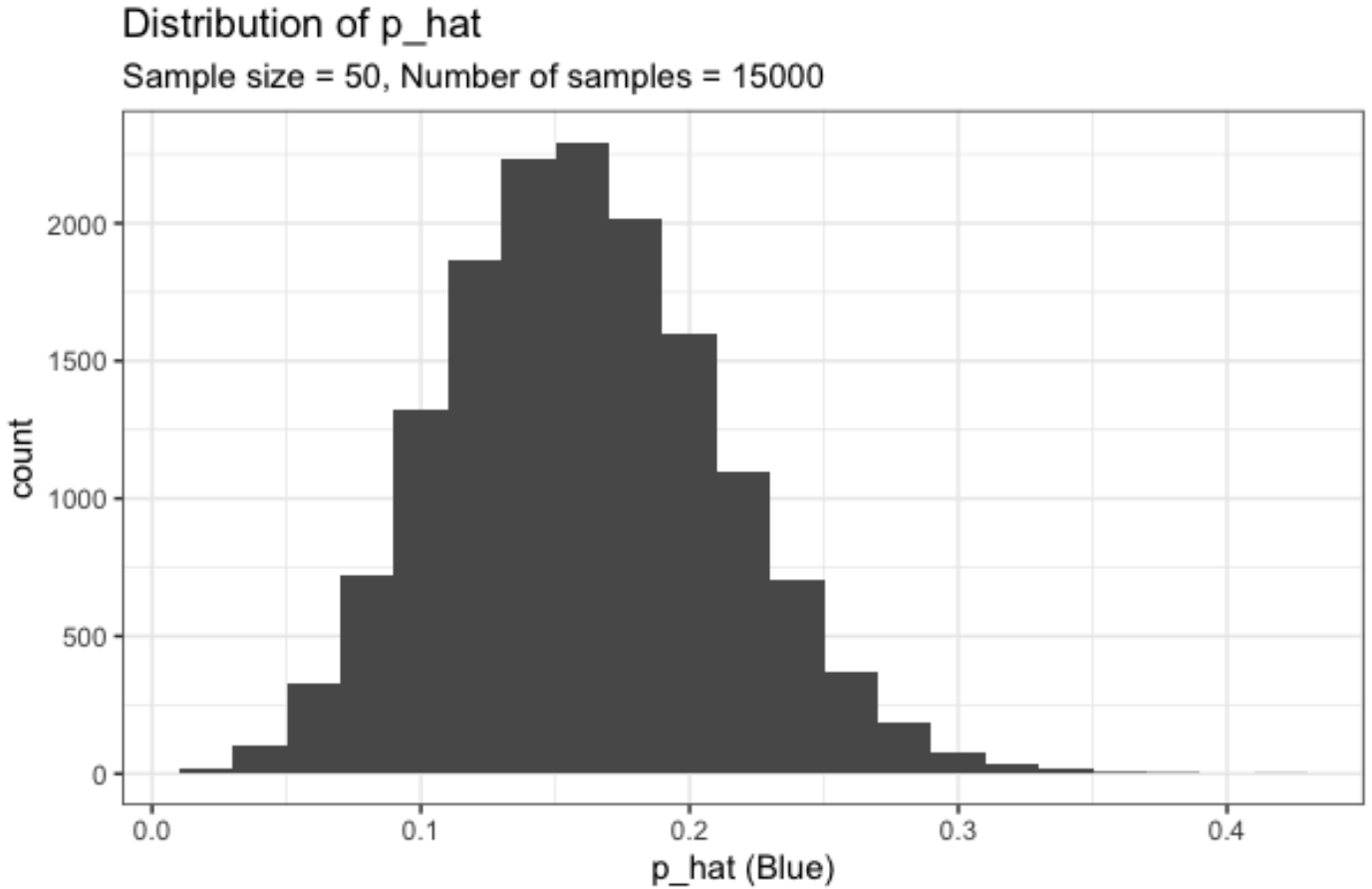
Based on this histogram, what's your estimate for the proportion of blue M&M's in the entire 100,000 M&M's gift?



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 50 M&M's at a time from the population

Remember our key question and answer: to assess uncertainty of a statistic, we need to know how much the statistic varies from sample to sample!

The variability of the sample statistic is so important that it gets it's own name...

The standard error of a statistic, SE, is the standard deviation of the sample statistic

The standard error can be calculated as the standard deviation of the sampling distribution

Poll question

The more the statistic varies from sample to sample, the standard error becomes

1. higher
2. lower

1. If you take random samples, the sampling distribution will be centered around the true population parameter
2. If sampling bias exists (if you do not take random samples), your sampling distribution may give you bad information about the true parameter

Outline

1. Today: sampling distribution

2. A sampling activity

3. Main take-away

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 $\rightarrow$ **Standard Error**