Notes 08 - Sampling Distributions and Bootstrap Resampling

STS 2300 (Fall 2024)

Updated: 2024-10-11

Table of Contents

[Reading for Notes 08 1](#_Toc179644039)

[Learning Goals for Notes 08 1](#_Toc179644040)

[Sampling Vocabulary Review 2](#_Toc179644041)

[House of Representatives Example 3](#_Toc179644042)

[A single sample of 30 seats 4](#_Toc179644043)

[Sampling Distributions 5](#_Toc179644044)

[Sample Size and Standard Error 7](#_Toc179644045)

[Bootstrap distributions 9](#_Toc179644046)

[Putting This Together 10](#_Toc179644047)

[Revisiting the Learning Goals for Notes 08 11](#_Toc179644048)

# Reading for Notes 08

Read [Chapter 7](https://moderndive.com/7-sampling.htm) of the Modern Dive textbook.

# Learning Goals for Notes 08

* Be able to describe the idea behind a sampling distribution and how it differs from the distribution of a single population or sample.
* Be able to describe the difference between standard deviations and standard errors.
* Understand how sample size relates to standard errors and sampling variability.
* Be able to describe how a bootstrap resampling distribution is used to mimic a sampling distribution and where the two distributions differ.

I’ll be using the following packages in this set of notes, so I’ll load them before I get started.

library(ggplot2)  
library(dplyr)  
library(moderndive)

## Warning: package 'moderndive' was built under R version 4.3.3

library(patchwork)

## Warning: package 'patchwork' was built under R version 4.3.3

# Sampling Vocabulary Review

So far this semester, we have largely focused on working with data and describing it both numerically and visually. However, the ultimate goal of statistics is usually to take our data/sample and to apply what we learn to a population. To do that, we need to have a better understanding of how sampling is done and how we can use a single sample to make conclusions about the larger group from which it came.

Let’s review some important statistics vocabulary you may have encountered in previous courses. Chapter 7 of the ModernDive textbook also covers most of these terms.

* The complete group of people / things we are interested in is called the **population**.
* A number describing / summarizing a population is known as a **parameter**. Examples include:
  + population proportions,
  + population means, (Greek letter mu)
  + population standard deviations, (Greek letter sigma)
* The group that we actually have information from is our **sample**.
* Numbers that summarize/describe a sample are called **statistics**. Examples include:
  + sample proportions, (p hat)
  + sample means, (x bar)
  + sample standard deviations,
* If our sample consists of the entire population, we call this a **census**. If we have a census, we can describe our sample like we’ve been doing. However, censuses are pretty rare.

When we don’t have a census, we want to make *inference* from our sample to our population. In other words, we want to take the results of our sample and *generalize* them to say something about our population. It only makes sense to do this if we have reason to believe our sample is *similar* to our population (i.e., has similar distributions of important variables), or in other words when our sample is a **representative sample**.

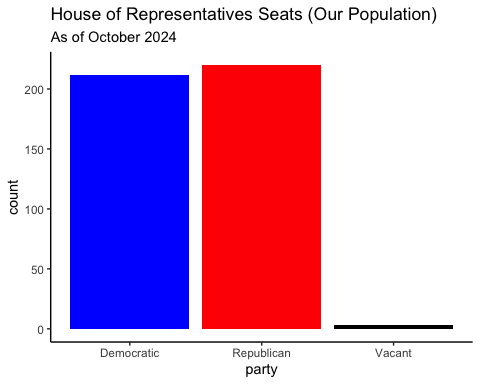
One great way to get a representative sample is to take a random sample (that is sufficiently large). A **random sample** is when we use some sort of chance device to make sure every member of the population has a chance of being selected for the sample. While taking a random sample doesn’t *guarantee* our sample is representative of our population, it does mean that we won’t consistently get samples that are **biased** and misrepresent our population in a particular way.

So… how does this all work? Let’s investigate!

# House of Representatives Example

The current House of Representatives has 212 Democratic members, 220 Republican members, and 3 vacant seats. Below is a visual of our population.

house\_of\_reps <- read.csv("https://raw.githubusercontent.com/nbussberg/STS2300-Fall2024/refs/heads/main/Data/house\_of\_reps.csv")  
  
pop\_dist <- ggplot(house\_of\_reps) +  
 geom\_bar(aes(x = party, fill = party),   
 show.legend = FALSE) +  
 theme\_classic() +  
 scale\_fill\_manual(values = c("blue", "red", "black"))   
  
pop\_dist + labs(title = "House of Representatives Seats (Our Population)",  
 subtitle = "As of October 2024")



This is the **population distribution** or **distribution of our population**. It tells us which values are possible and how frequent they are *in the entire group we are interested in*. However, we normally don’t have access to this whole group, so instead we take a sample.

## A single sample of 30 seats

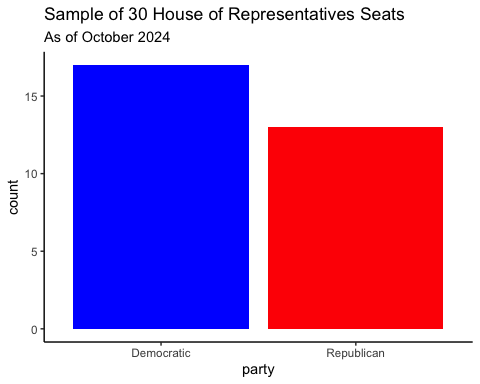
Suppose I don’t know the party breakdown of the House of Representatives, but I’m curious what proportion of the seats in the House of Representatives belong to the Democratic Party.

mysamp <- rep\_sample\_n(house\_of\_reps, size = 30)  
table(mysamp$party)

##   
## Democratic Republican   
## 17 13

Above I’ve taken a simple random sample of 30 seats. I’m using the rep\_sample\_n() function (from the moderndive package associated with our textbook) to randomly choose 30 rows from a data frame I’ve created called house\_of\_reps. Then I’m storing it in a new data frame called mysamp. Using the table() function I can see there are 17 Democrats and 13 Republicans in my sample (and 0 vacant seats).

sample\_dist <- ggplot(mysamp) +  
 geom\_bar(aes(x = party, fill = party),   
 show.legend = FALSE) +  
 theme\_classic() +  
 scale\_fill\_manual(values = c("blue", "red", "black"))  
  
sample\_dist + labs(title = "Sample of 30 House of Representatives Seats",  
 subtitle = "As of October 2024")



The graph above is the **distribution of my sample** (note that this is **not** the same as a sampling distribution). You can see that 17 of the 30 seats in my sample belong to the Democratic Party. Thus, the sample proportion (a statistic) is = 17 / 30 = 0.567. In other words, 56.7% of the seats in my sample belong to the Democratic Party.

**Question:** Will you also get an estimate of 56.7% when you run this code? Why or why not? Try it out after you answer.

**Answer:** Everyone might get slightly different answers because each sample is random. This means that we will get different representatives in each sample.

If this is all the information I have, my best guess would be that 56.7% of all the seats belong to the Democratic party. Is my guess exactly correct? Probably not, but I hope it’s “close”. Let’s see if we can figure out about how close it is.

**Note:** When using randomization in our code, it’s important that we **set a “seed”** so that we get the same results the next time we run our code (or when someone else runs our code). We can do this with the set.seed() function.

**Example** Go back to before where you used the rep\_sample\_n() function and add a line that reads set.seed(82720). Then re-run your code. What happens?

**Answer:** When we run the set.seed() function first, we get the same sample that was in the notes.

# Sampling Distributions

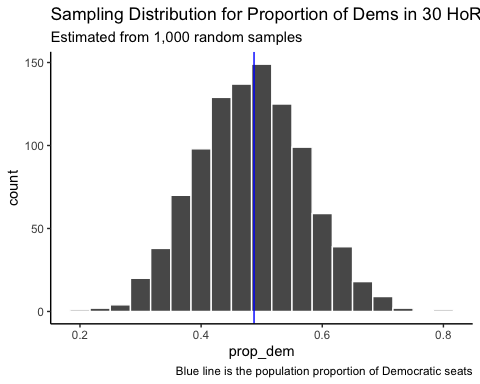
A **sampling distribution** describes the possible values for a *statistic* over all possible random samples of a certain size.

Let’s try to simulate the sampling distribution for the proportion of Democratic members in random samples of 30 seats from the House of Representatives.

The rep\_sample\_n() function has an argument called reps that will allow us to specify how many random samples we want to take. Let’s take 1,000 random samples of 30 seats and compare the sample proportions that we get in these samples.

The rep\_sample\_n() function will give us the raw data for all 1,000 random samples. Thus, it contains 30,000 rows (1,000 samples \* 30 observations per sample). If we want to look at our sample proportions, we can use summarize(). Normally, I would use the .by argument, but rep\_sample\_n() produces a “grouped data frame” that is already grouped by replicate (aka sample).

true\_p <- mean(house\_of\_reps$party == "Democratic")  
  
my\_samples\_n30 <- house\_of\_reps %>%  
 rep\_sample\_n(size = 30,  
 reps = 1000)  
  
my\_phats\_n30 <- my\_samples\_n30 %>%  
 summarize(prop\_dem = mean(party == "Democratic"))  
  
sampling\_dist <- ggplot(my\_phats\_n30) +  
 geom\_histogram(aes(x = prop\_dem),  
 binwidth = 1/30,  
 color = "white") +  
 theme\_classic() +  
 scale\_fill\_manual(values = c("blue", "red", "black"))  
  
sampling\_dist +   
 labs(title = "Sampling Distribution for Proportion of Dems in 30 HoR Seats",  
 subtitle = "Estimated from 1,000 random samples",  
 caption = "Blue line is the population proportion of Democratic seats") +  
 geom\_vline(xintercept = true\_p,  
 color = "blue")



The graph above represents our *estimated sampling distribution* of the proportion of House of Representative seats that belong to the Democratic party. It is *estimated* because we looked at 1,000 random samples, not all random samples. Your graph may look slightly different from mine if you used a different seed.

What can we learn from this graph?

**Answer:** On average, we are getting sample proportions around the population proportion. However, we are not always getting super close to the population proportion. In the sampling distribution above, we see sample proportions anywhere from 0.2 up to almost 0.8. This graph seems to be in a Normal distribution or bell-shaped pattern.

A **standard error** is the standard deviation in the estimates you get from repeated random samples. In other words, it tells us how much our estimates should vary from sample to sample. This is useful for us to know roughly how accurate of an estimate we might have.

We can see what the estimated standard error is for our estimated sampling distribution by calculating the standard deviation of our 1,000 sample proportions we found above.

sd(my\_phats\_n30$prop\_dem)

## [1] 0.08886335

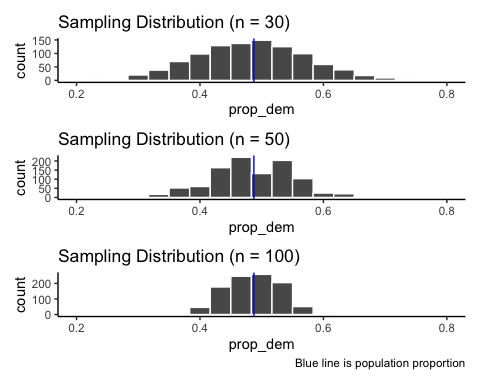
The standard error for samples of 30 seats is around 0.089. (Note: You may have gotten a different number if you didn’t include a seed or used a different seed than me.)

**Question:** What will happen to the standard error if I take random samples of 50 or 100 seats?

**Answer:** As our sample size increases, we end up with better estimates of our parameter (population proportion). This means that the standard error gets smaller because our estimates tend to be closer to the true value for the population than when we have a smaller sample size.

# Sample Size and Standard Error

Let’s take 1,000 random samples of 50 and 100 seats and calculate the proportion of seats belonging to the Democratic Party for both. Then let’s compare them to what we saw when we took samples of 30 seats.



Notice how each of the sampling distributions is centered on the population proportion and how the sampling distribution narrows as the sample size increases. This tells us that as we take bigger random samples, we are less likely to get estimates far away from our parameter of interest.

We can compare the standard errors as well.

se30 <- sd(my\_phats\_n30$prop\_dem)  
se50 <- sd(my\_phats\_n50$prop\_dem)  
se100 <- sd(my\_phats\_n100$prop\_dem)  
  
data.frame(n = c(30, 50, 100),   
 SE = round(c(se30, se50, se100), 3))

## n SE  
## 1 30 0.089  
## 2 50 0.065  
## 3 100 0.044

With samples of 30 seats, it is typical that our estimate may be off by around 0.089.

With samples of 50 seats, it is typical that our estimate may be off by around 0.065.

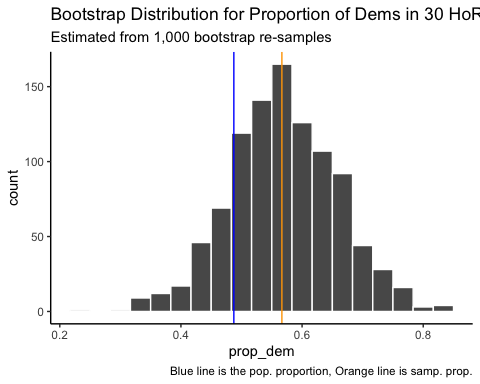
With samples of 100 seats, it is typical that our estimate may be off by around 0.044.

# Bootstrap distributions

So far we have talked about the distribution of a population, the distribution of a sample, and the sampling distribution. However, only one of those (the distribution of our sample) is something we usually have access to in real life. The other two distributions require us to have access to the entire population, which is unrealistic in most cases.

A bootstrap distribution is a way for us to mimic a sampling distribution using only our sample. To do this, we recreate the process of taking many samples, but we take them from our original sample (which we have access to) instead of from the population (which we don’t). For this to work, we have to take our samples **with replacement**. Let’s see what happens when we do this.

phat <- mean(mysamp$party == "Democratic")  
  
myboot <- mysamp %>%  
 ungroup() %>%  
 select(-replicate) %>%  
 rep\_sample\_n(size = 30,  
 reps = 1000,  
 replace = TRUE) %>%  
 summarize(prop\_dem = mean(party == "Democratic"))  
  
boot\_dist <- ggplot(myboot) +  
 geom\_histogram(aes(x = prop\_dem),  
 binwidth = 1/30,  
 color = "white") +  
 theme\_classic() +  
 scale\_fill\_manual(values = c("blue", "red", "black")) +  
 geom\_vline(xintercept = true\_p,  
 color = "blue") +  
 geom\_vline(xintercept = phat,  
 color = "orange")  
  
boot\_dist +  
 labs(title = "Bootstrap Distribution for Proportion of Dems in 30 HoR Seats",  
 subtitle = "Estimated from 1,000 bootstrap re-samples",  
 caption = "Blue line is the pop. proportion, Orange line is samp. prop.")



This should look a lot like the sampling distribution, but it will be centered on my sample proportion instead of on the population proportion. In the next set of notes, we will see how this will allow us to create confidence intervals for any parameter of interest!

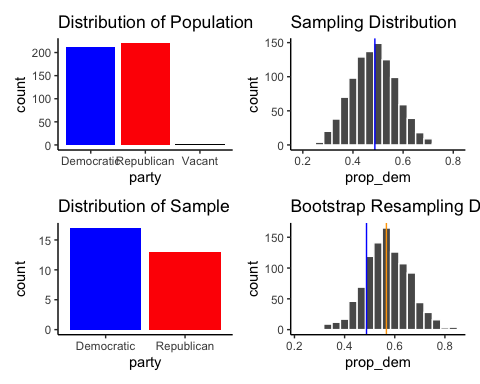
# Putting This Together

The distribution of a population and distribution of a sample both describe distributions of individual observations. We can think of the distribution of a sample as being our best guess at what the population generally looks like.

The sampling distribution and bootstrap distribution both describe distributions of statistics for samples of a certain size. We can think of the bootstrap distribution as being our best guess for what the sampling distribution looks like.

The distribution of the population and sampling distribution will both be “centered” on the parameter of interest. The distribution of the sample and the bootstrap distribution will both be “centered” on the statistic we observed in our sample.

We can see a visual of this below:



# Revisiting the Learning Goals for Notes 08

* Be able to describe the idea behind a sampling distribution and how it differs from the distribution of a single population or sample.
* Be able to describe the difference between standard deviations and standard errors.
* Understand how sample size relates to standard errors and sampling variability.
* Be able to describe how a bootstrap resampling distribution is used to mimic a sampling distribution and where the two distributions differ.