Data606 Lab 5a - Sampling distributions

Avery Davidowitz

2022-10-16

```
knitr::opts_chunk$set(echo = TRUE)
```

Load Packages and set seed

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

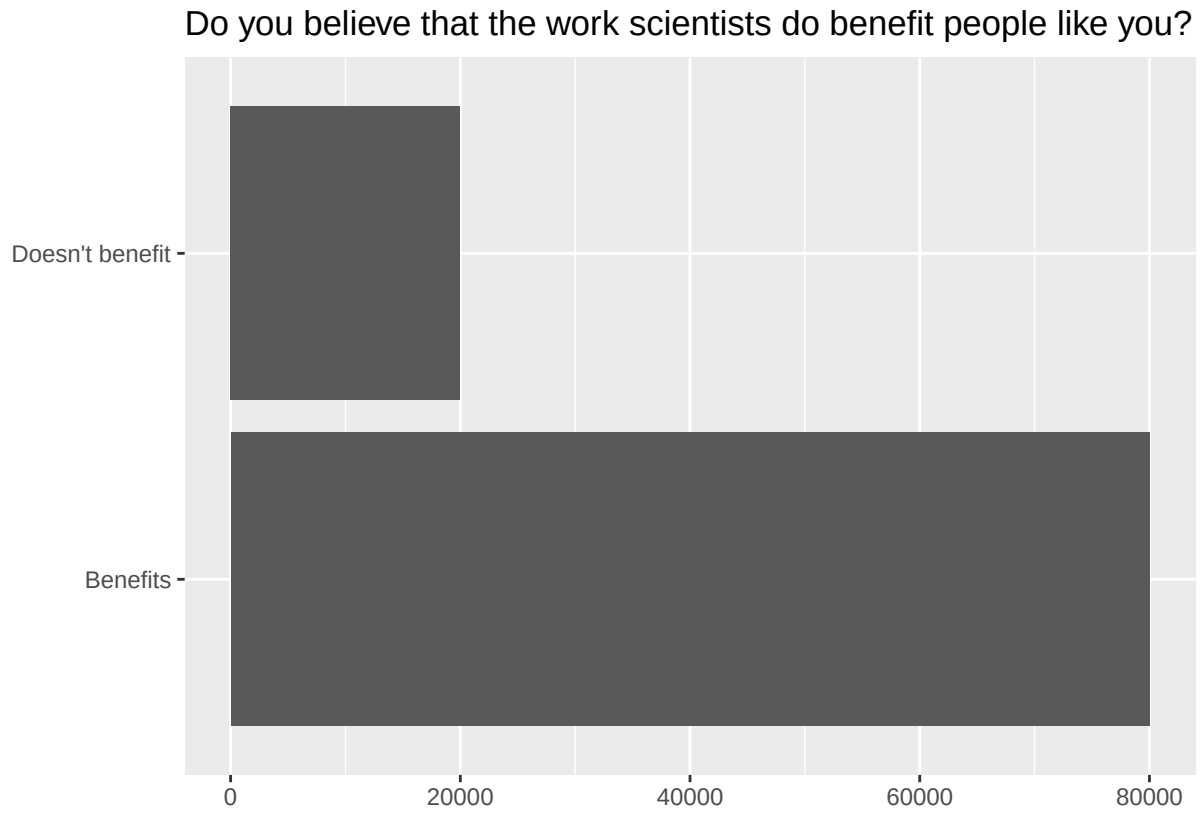
```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
set.seed(86)
```

Population

```
global_monitor <- tibble(
  scientist_work = c(rep("Benefits", 80000), rep("Doesn't benefit", 20000))
)
ggplot(global_monitor, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```




```
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits        80000   0.8
## 2 Doesn't benefit 20000   0.2
```

Exercise 1

Describe the distribution of responses in this sample. How does it compare to the distribution of responses in the population.

```
samp1 <- global_monitor %>%
  sample_n(50)
samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
```

```
## 1 Benefits           40   0.8
## 2 Doesn't benefit    10   0.2
```

The sample exactly equals the proportions of the population.

Exercise 2

Would you expect the sample proportion to match the sample proportion of another student's sample? Why, or why not? If the answer is no, would you expect the proportions to be somewhat different or very different? Ask a student team to confirm your answer.

I would expect samples to be very close to the population proportion and my sample. The central limit theorem applies to this sample since the sample is random and therefore independent and since np = 40 and n(1-p) = 10 both greater or equal to 10. Therefore, the standard error for a sample proportion is equal to:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.8 * .2}{50}} =$$

```
SE_phat <- (.8 * .2 / 50) ** (1/2)
print(SE_phat * 1.96)
```

```
## [1] 0.1108743
```

Since the standard error times Z for a 95% confidence interval is less than 1, I would expect 95% of students to have between 39 and 41 Benefits in their samples.

Exercise 3

Take a second sample, also of size 50, and call it samp2. How does the sample proportion of samp2 compare with that of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population proportion?

```
samp2 <- global_monitor %>%
  sample_n(50)
samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

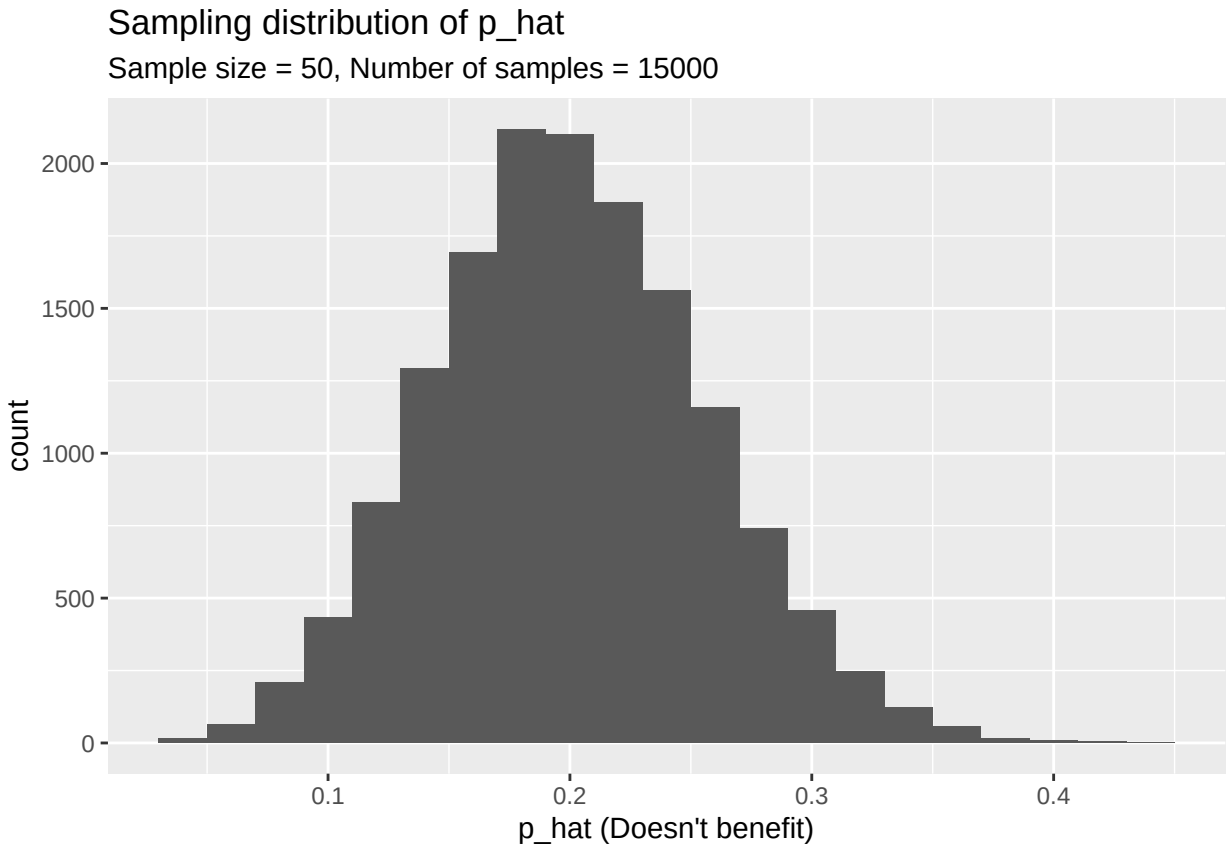
```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           37  0.74
## 2 Doesn't benefit    13  0.26
```

The second sample is substantially different with a p hat of .74. As the sample sizes increase the error between p hat and p will decrease due to n being in the denominator of the standard error. A sample of 1000 will likely be more accurate than 100.

Exercise 4

How many elements are there in sample_props50? Describe the sampling distribution, and be sure to specifically note its center. Make sure to include a plot of the distribution in your answer.

```
sample_props50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```



sample_props50 contains 15,000 rows of observations of simulated samples. It is nearly normally distributed and centered around the .2 value for 1-p of the overall population.

Exercise 5

To make sure you understand how sampling distributions are built, and exactly what the rep_sample_n function does, try modifying the code to create a sampling distribution of 25 sample proportions from

samples of size 10, and put them in a data frame named sample_props_small. Print the output. How many observations are there in this object called sample_props_small? What does each observation represent?

```
sample_props_small <- global_monitor %>%
                    rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
print(sample_props_small)
```

```
## # A tibble: 23 x 4
## # Groups:   replicate [23]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     2   0.2
##  2         2 Doesn't benefit     2   0.2
##  3         3 Doesn't benefit     5   0.5
##  4         4 Doesn't benefit     3   0.3
##  5         5 Doesn't benefit     5   0.5
##  6         7 Doesn't benefit     3   0.3
##  7         8 Doesn't benefit     1   0.1
##  8         9 Doesn't benefit     2   0.2
##  9        10 Doesn't benefit     3   0.3
## 10        11 Doesn't benefit     1   0.1
## # ... with 13 more rows
```

The code represents a sequence of 25 samples with sample size n = 10 from the generated global monitor set. This data frame only has 23 rows since we have 2 samples that had 0 Doesn't benefit values. This sample strategy is flawed and therefore not accurately representing the population. We are violating the success-failure condition needed to apply the central limit theorem.

Exercise 6

Use the app below to create sampling distributions of proportions of Doesn't benefit from samples of size 10, 50, and 100. Use 5,000 simulations. What does each observation in the sampling distribution represent? How does the mean, standard error, and shape of the sampling distribution change as the sample size increases? How (if at all) do these values change if you increase the number of simulations? (You do not need to include plots in your answer.)

```
sample_props10 <- global_monitor %>%
                    rep_sample_n(size = 10, reps = 5000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
se_10 <- (.8 * .2 / 10) ** (1/2)
mean_10 <- sample_props10 %>%
  group_by(scientist_work) %>%
  dplyr::summarise(avg = mean(p_hat))
print(se_10)
```

```
## [1] 0.1264911
```

```
print(mean_10)
```

```
## # A tibble: 1 x 2
##   scientist_work    avg
##   <chr>           <dbl>
## 1 Doesn't benefit 0.222
```

```
sample_props50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 5000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
se_50 <- (.8 * .2 / 50) ** (1/2)
mean_50 <- sample_props50 %>%
  group_by(scientist_work) %>%
  dplyr::summarise(avg = mean(p_hat))
print(se_50)
```

```
## [1] 0.05656854
```

```
print(mean_50)
```

```
## # A tibble: 1 x 2
##   scientist_work    avg
##   <chr>           <dbl>
## 1 Doesn't benefit 0.201
```

```
sample_props100 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 5000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
se_100 <- (.8 * .2 / 100) ** (1/2)
mean_100 <- sample_props100 %>%
  group_by(scientist_work) %>%
  dplyr::summarise(avg = mean(p_hat))
print(se_100)
```

```
## [1] 0.04
```

```
print(mean_100)
```

```
## # A tibble: 1 x 2
##   scientist_work    avg
##   <chr>           <dbl>
## 1 Doesn't benefit 0.200
```

As the sample size increases the standard error decreases and the sample mean moves closer to the true population mean of .2.

Exercise 7

Take a sample of size 15 from the population and calculate the proportion of people in this sample who think the work scientists do enhances their lives. Using this sample, what is your best point estimate of the population proportion of people who think the work scientists do enhances their lives?

```
samp3 <- global_monitor %>%
  sample_n(15)
samp3 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

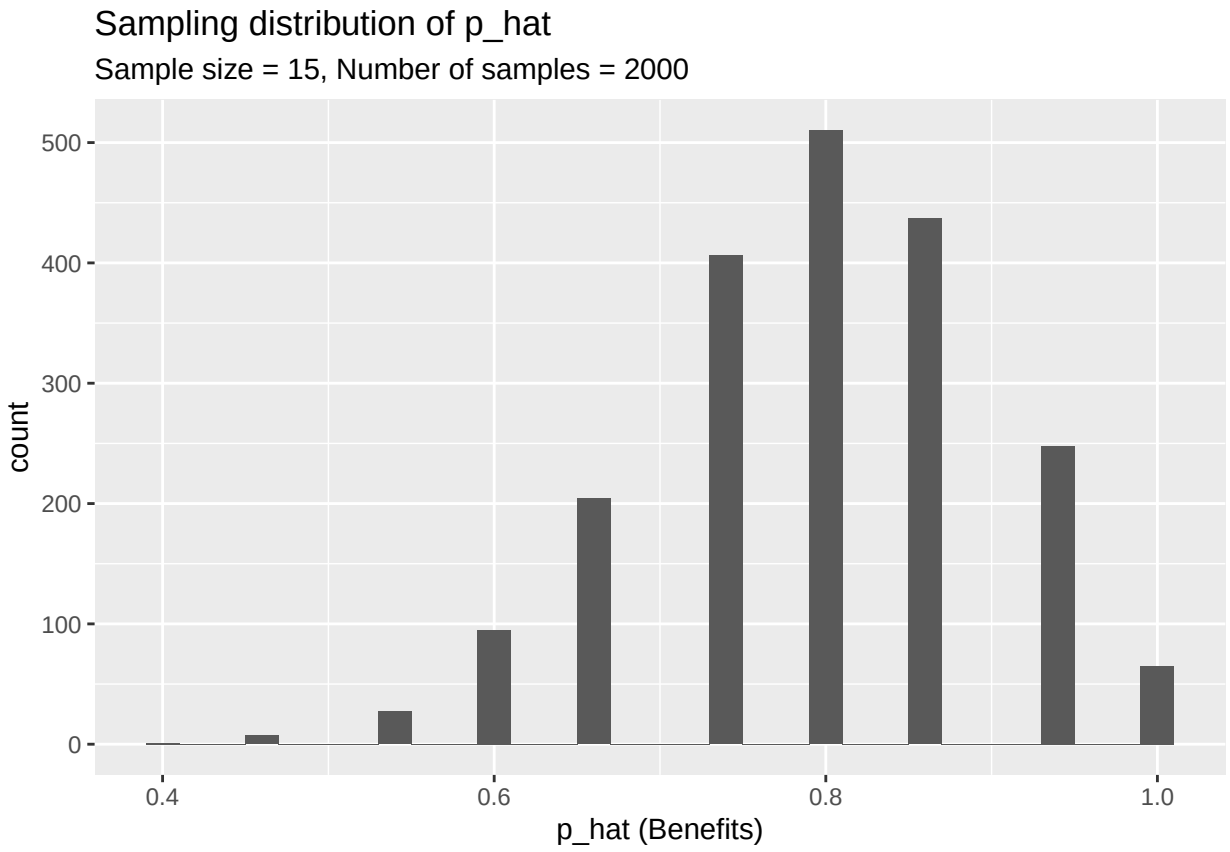
```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           12   0.8
## 2 Doesn't benefit     3   0.2
```

My best estimate is 12/15 or .8.

Exercise 8

Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enhance their lives for samples of size 15 by taking 2000 samples from the population of size 15 and computing 2000 sample proportions. Store these proportions in as sample_props15. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the true proportion of those who think the work scientists do enhances their lives to be? Finally, calculate and report the population proportion.

```
sample_props15 <- global_monitor %>%
                    rep_sample_n(size = 15, reps = 2000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")
ggplot(data = sample_props15, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefits)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 15, Number of samples = 2000"
  )
```

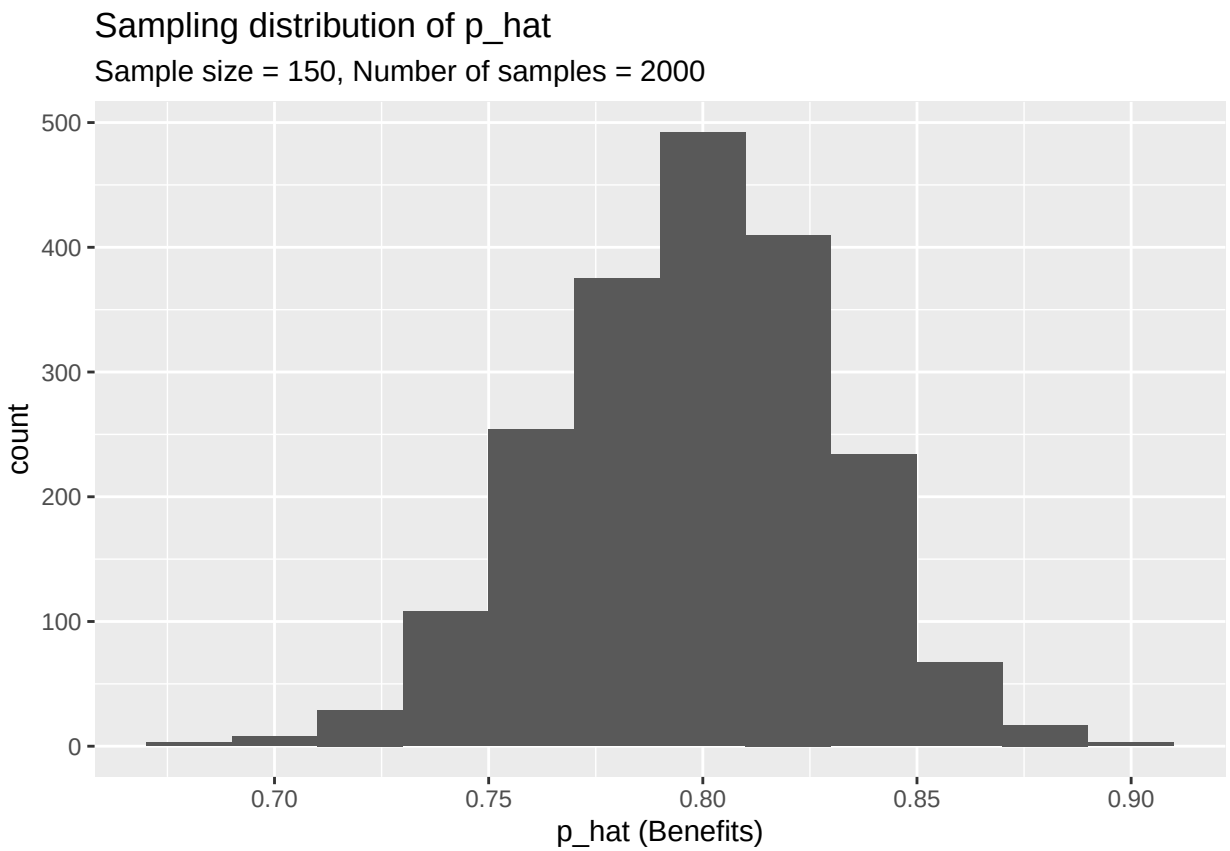


Based on the center of this nearly normal distribution of samples I would estimate the true proportion to be .8 which is equal to the population proportion.

Exercise 9

Change your sample size from 15 to 150, then compute the sampling distribution using the same method as above, and store these proportions in a new object called sample_props150. Describe the shape of this sampling distribution and compare it to the sampling distribution for a sample size of 15. Based on this sampling distribution, what would you guess to be the true proportion of those who think the work scientists do enhances their lives?

```
sample_props150 <- global_monitor %>%
                    rep_sample_n(size = 150, reps = 2000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")
ggplot(data = sample_props150, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefits)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 150, Number of samples = 2000"
  )
```



The distribution for the samples of size 150 is very similar to the samples of size 15 with a nearly normal shape and centered around the true population mean of .8. However, the 15 and 150 vary significantly by the spread.

Exercise 10

Of the sampling distributions from 2 and 3, which has a smaller spread? If you're concerned with making estimates that are more often close to the true value, would you prefer a sampling distribution with a large or small spread?

The samples of size 150 had significantly less spread. If I was concerned with making estimates that were close to the true population value, I would choose a large sample size.