Data606 - Lab 5A

Cameron Smith

2020-10-05

```
library(tidyverse)
library(openintro)
library(infer)
library(psych)
```

The unknown sampling distribution

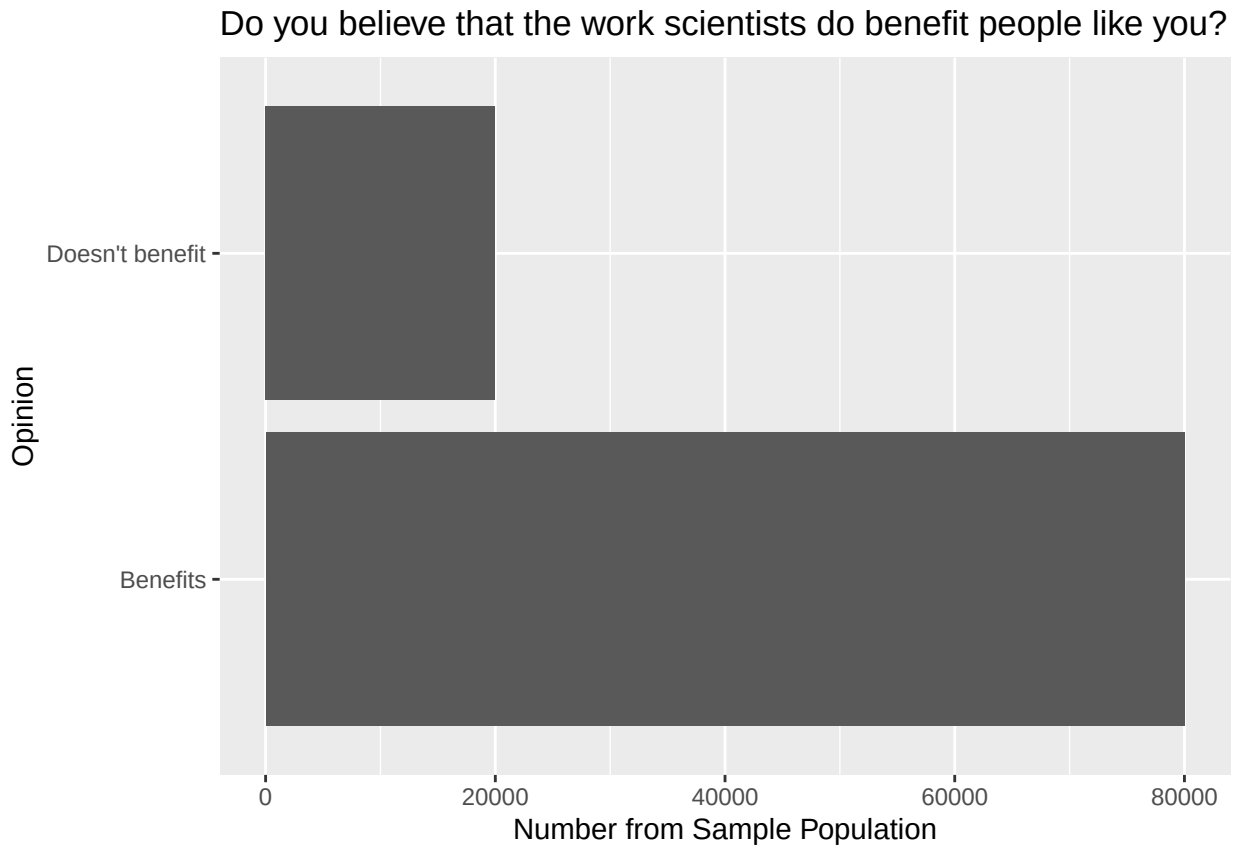
Exercise 1

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

The distribution of the sample population is pretty different from the that of the total population. In particular the "Benefits" opinion has increased from 80% to 88% and, inversely, the "Doesn't benefit" opinion has decreased from 20% to 12%.

```
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 = "Opinion", y = "Number from Sample Population",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```



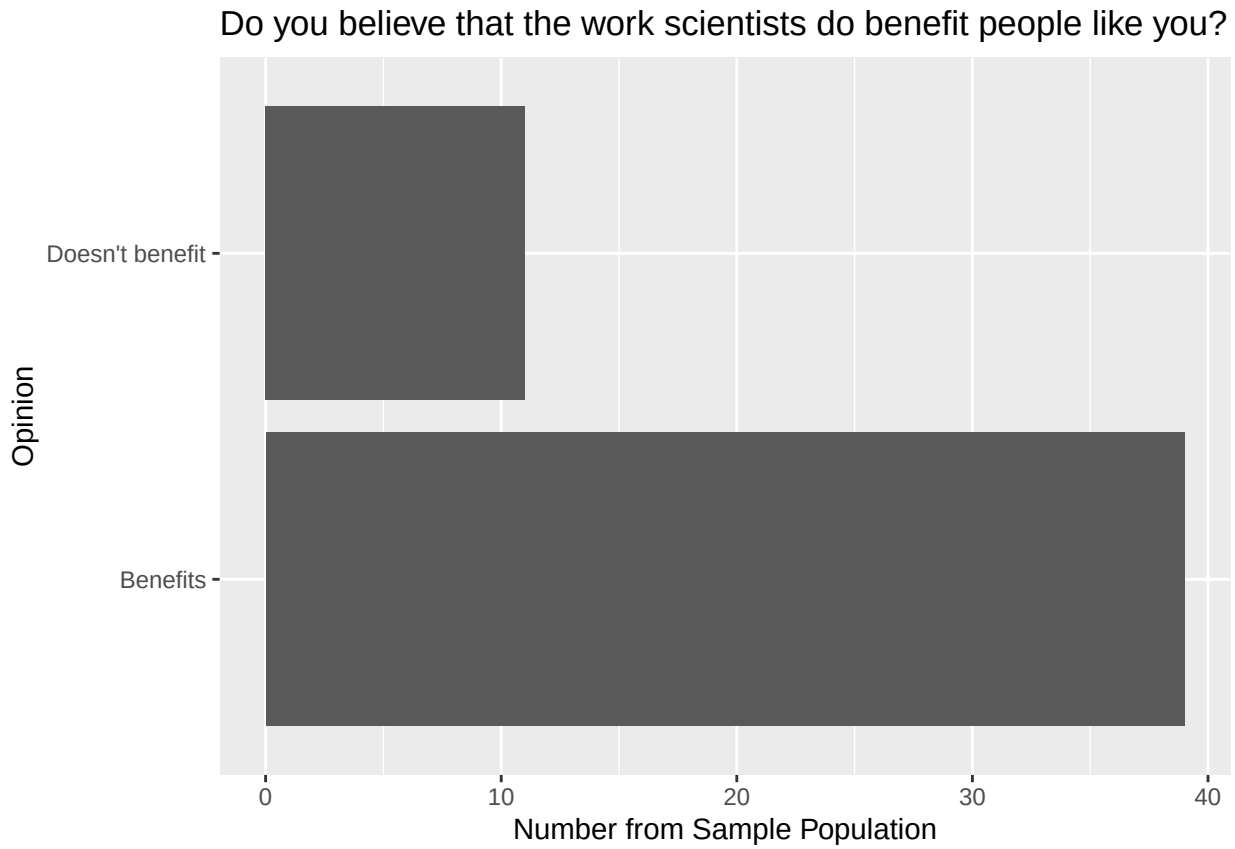

```
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

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

```
samp1 <- global_monitor %>%
  sample_n(50)

# Code for the exercise

ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "Opinion", y = "Number from Sample Population",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```




```
# Summary statistics
samp1 %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           39  0.78
## 2 Doesn't benefit    11  0.22
```

Exercise 2

Would you expect the sample proportion to match the sample proportion of another student's sample? Why, or why not?

No, I would not expect another student's sample to perfectly match mine, although I would expect it to be close. The reason for this is that the random seed used to select the sample population would be different (unless intentionally set to be the same). We can see how this works in the example below. Each time the loop runs through the same code the figures change a bit.

```
for (i in 1:3) {
global_monitor %>% sample_n(50) %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n)) %>% print()
}
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           36  0.72
## 2 Doesn't benefit    14  0.28
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           38  0.76
## 2 Doesn't benefit    12  0.24
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           41  0.82
## 2 Doesn't benefit     9  0.18
```

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?

As illustrated above (in the code I used fr example 2), the sample proportion will change each time unless the random seed is set. If we look at samples of 100 and 1000 then I would expect that 1000 would have a more accurate estimate of the population proportation. This is based on the law of large numbers.

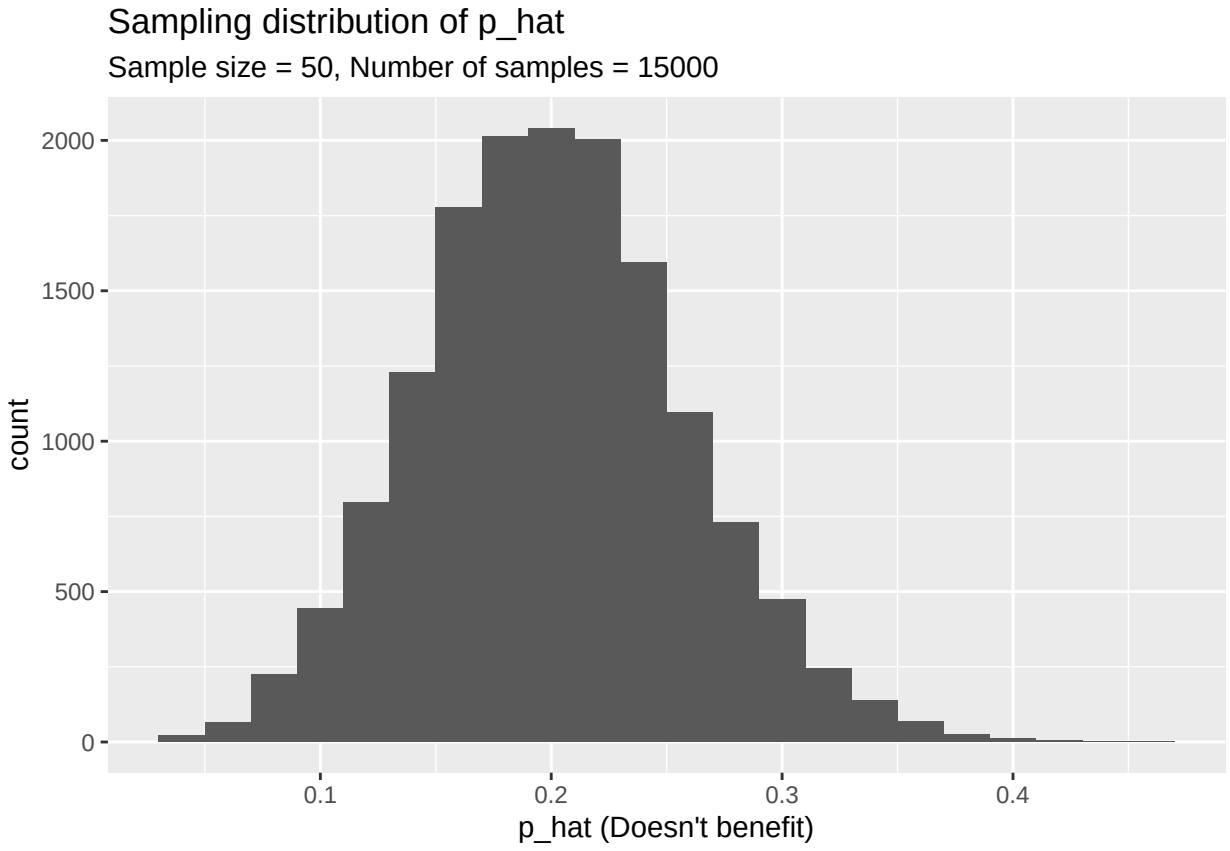
Exercise 4

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.

There are 15,000 elements in sample_props 50. The distribution is normal, and bimodal based on the below histogram although if we adjust the bin width then it would be unimodal. Interestingly the center is approximately 20%, which is the same as that of the total population.

```
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"
  )
```




```
# Code for exercise

nrow(sample_props50)
```

```
## [1] 15000
```

Interlude: Sampling distributions

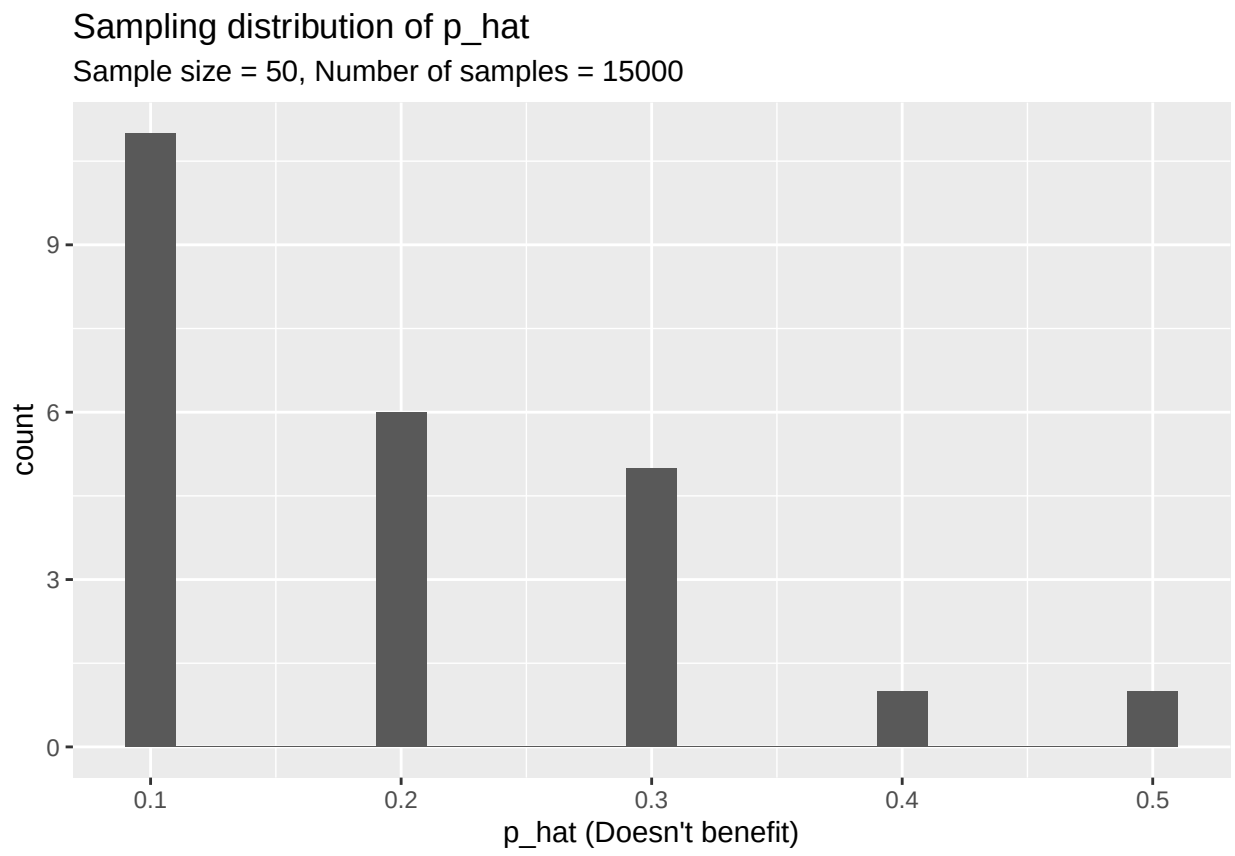
Exercise 1

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?

There are 25 observations in sample_props_small, and each of them represents a different sample taken via the rep_sample_n function; specifically the proportion of the sample population that think the work scientists do doesn't benefit them.

```
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")

ggplot(data = sample_props_small, 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 size and the sampling distribution

Exercise 1

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, standar 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.)

The sample population figures more closely reflect those of the total population the higher the sample population size and the number of times that the sample is run. This is illustrated in the examples below, and we can see that when the number of repetitions gets large (1000, 10000, 100000) the mean reflects that of the total population, and the standard error is reduced to zero.

Each observation represents a different sample taken via the rep_sample_n function; specifically the proportion of the sample population that think the work scientists do doesn't benefit them (similar to Exercise 5 above).

Note: the lab refers to an app to use to answer this question but there was no app actually references in the text so I ran the calculations manually in the code block below.

```
sample_props_1 <- 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")

sample_props_2 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

sample_props_3 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

sample_props_big1 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 1000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

sample_props_big2 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 10000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

sample_props_big3 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 100000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
```

```
                  filter(scientist_work == "Doesn't benefit")

describe(sample_props_1$p_hat)
```

```
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 24 0.23 0.13    0.2    0.22 0.15 0.1 0.5   0.4 0.94    -0.06 0.03
```

```
describe(sample_props_2$p_hat)
```

```
##    vars  n mean   sd median trimmed  mad  min max range  skew kurtosis   se
## X1    1 25 0.21 0.06    0.2    0.21 0.06 0.08 0.3  0.22 -0.34    -0.65 0.01
```

```
describe(sample_props_3$p_hat)
```

```
##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 25 0.18 0.03   0.19    0.18 0.03 0.11 0.24  0.13 -0.46    -0.35 0.01
```

```
describe(sample_props_big1$p_hat)
```

```
##    vars    n mean   sd median trimmed  mad min  max range skew kurtosis se
## X1    1 1000  0.2 0.04    0.2     0.2 0.04 0.1 0.38  0.28 0.35     0.47  0
```

```
describe(sample_props_big2$p_hat)
```

```
##    vars     n mean   sd median trimmed  mad  min  max range skew kurtosis se
## X1    1 10000  0.2 0.04    0.2     0.2 0.04 0.08 0.35  0.27 0.11    -0.08  0
```

```
describe(sample_props_big3$p_hat)
```

```
##    vars     n mean   sd median trimmed  mad  min  max range skew kurtosis se
## X1    1 1e+05  0.2 0.04    0.2     0.2 0.04 0.03 0.39  0.36 0.15    -0.02  0
```

More Practice

Exercise 1

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

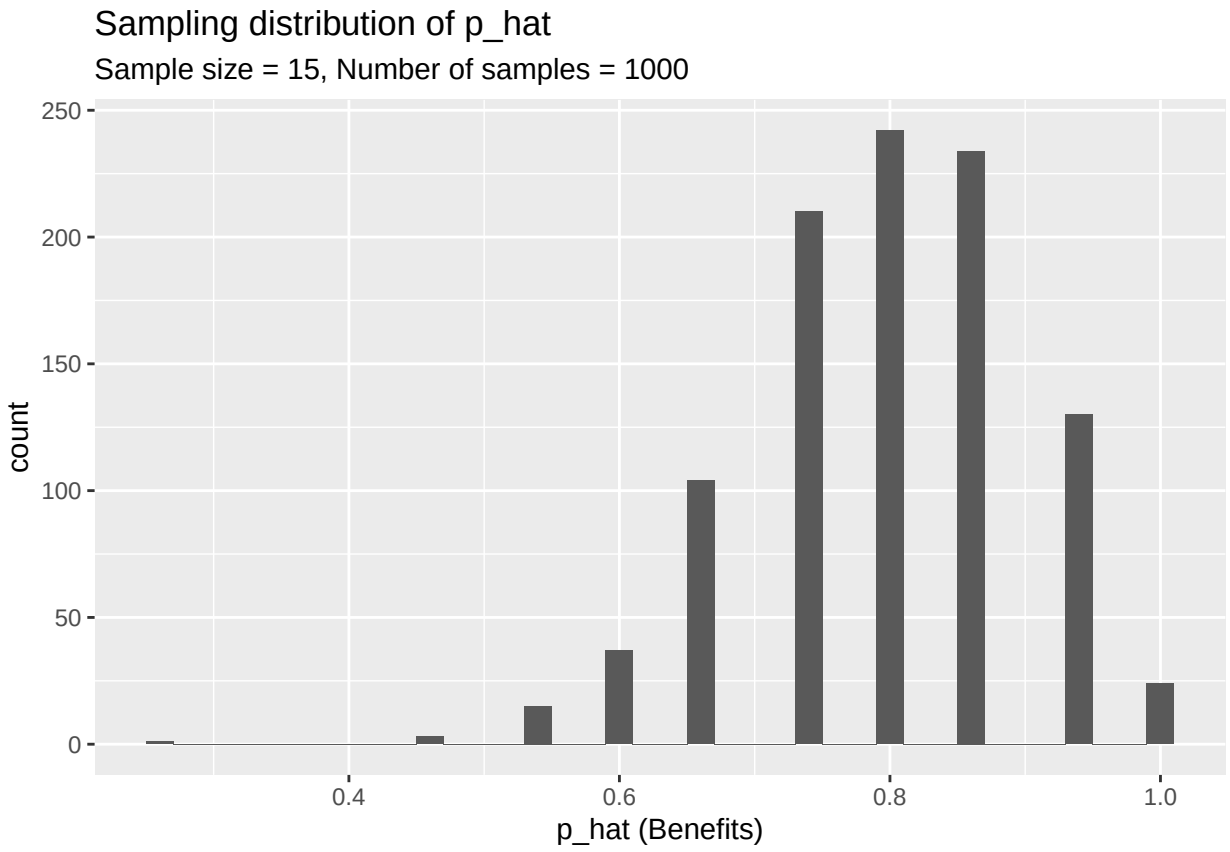
Based on the data and boxplot below we can see that the proportion of people who think the work that scientists do enhances their lives is 80% (this is the mean and median, and the standard error is 0).

```
sample_props_15 <- global_monitor %>%
                    rep_sample_n(size = 15, reps = 1000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")

describe(sample_props_15$p_hat)
```

```
##    vars    n mean  sd median trimmed mad  min max range  skew kurtosis se
## X1    1 1000  0.8 0.1    0.8     0.8 0.1 0.27   1  0.73 -0.47     0.49  0
```

```
ggplot(data = sample_props_15, 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 = 1000"
  )
```


Sampling distribution of p_hat
Sample size = 15, Number of samples = 1000
250
200
150
100
50
0
count
0.4
0.6
0.8
1.0
p_hat (Benefits)

Exercise 2

Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enchances 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 4 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 enchances their lives to be? Finally, calculate and report the population proportion.

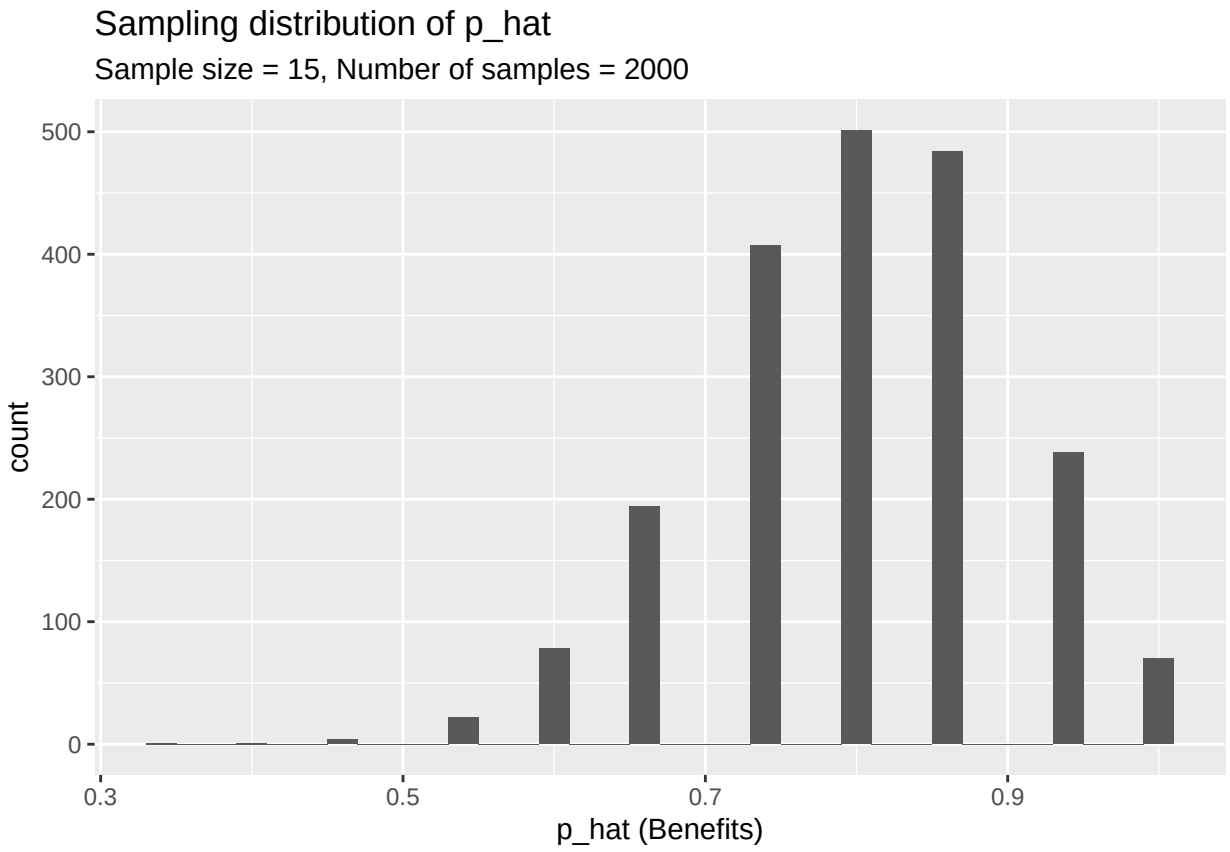
The distribution is normal, bimodal, and left-skewed. Based solely on looking at the histogram below, I would estimate the proportion of people who think the work that scientists do enhances their lives to be between 80% and 90%. The actual proportion is 80% as calculated using the describe() function below.

```
sample_props_15_B <- global_monitor %>%
                     rep_sample_n(size = 15, reps = 2000, replace = TRUE) %>%
                     count(scientist_work) %>%
                     mutate(p_hat = n /sum(n)) %>%
                     filter(scientist_work == "Benefits")

describe(sample_props_15$p_hat)
```

```
##    vars    n mean  sd median trimmed mad  min max range  skew kurtosis se
## X1    1 1000  0.8 0.1    0.8     0.8 0.1 0.27   1  0.73 -0.47     0.49  0
```

```
ggplot(data = sample_props_15_B, 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"
  )
```


Sampling distribution of p_hat
Sample size = 15, Number of samples = 2000
count
500
400
300
200
100
0
0.3
0.5
0.7
0.9
p_hat (Benefits)

Exercise 3

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 enchances their lives?

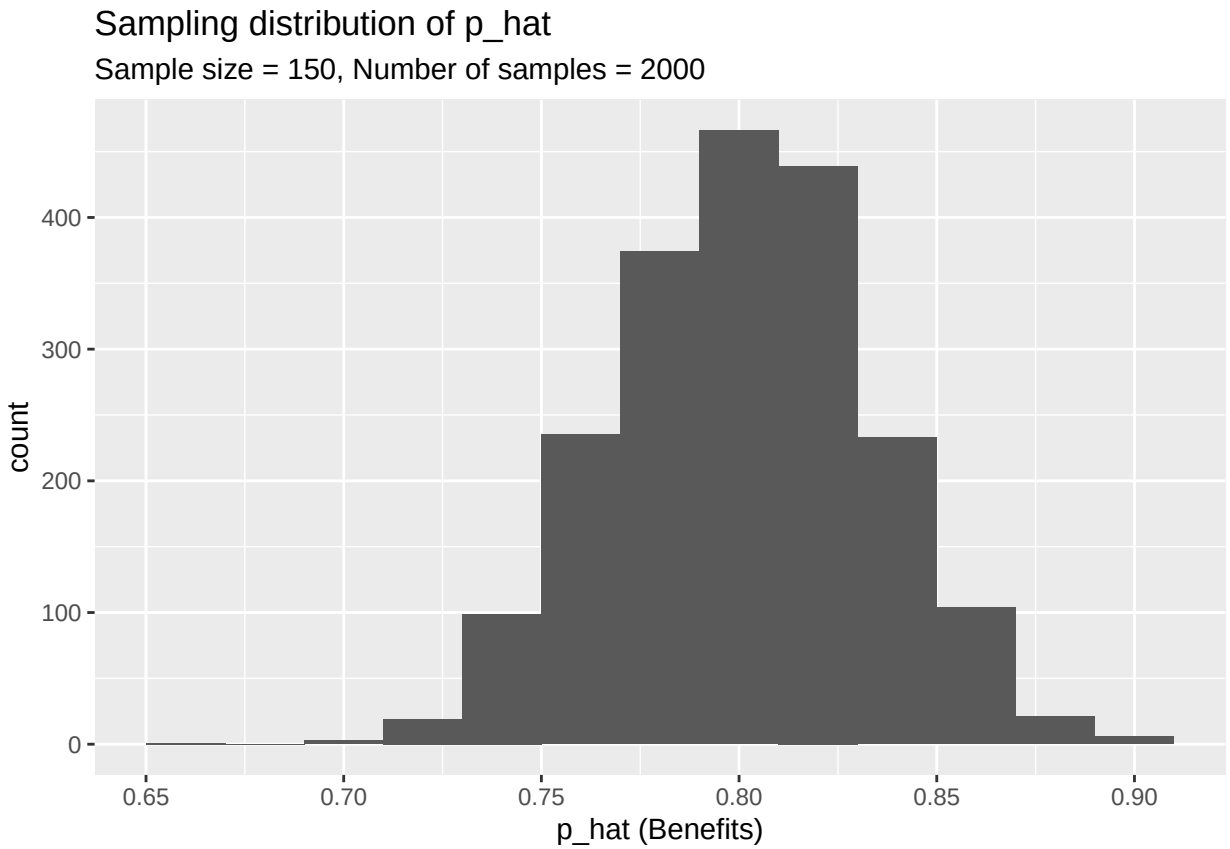
Based solely on looking at the histogram below, I would estimate the proportion of people who think the work that scientists do enhances their lives to be 80%. Unlike the previous example with a bimodal distribution, this is unimodal so the answer is pretty apparent.

```
sample_props_150 <- global_monitor %>%
                    rep_sample_n(size = 150, reps = 2000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")

describe(sample_props_150$p_hat)
```

```
##    vars    n mean   sd median trimmed  mad  min max range  skew kurtosis se
## X1    1 2000  0.8 0.03    0.8     0.8 0.03 0.66 0.9  0.24 -0.06    -0.09  0
```

```
ggplot(data = sample_props_150, 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"
  )
```


Sampling distribution of p_hat
Sample size = 150, Number of samples = 2000
count
400
300
200
100
0
0.65
0.70
0.75
0.80
0.85
0.90
p_hat (Benefits)

Exercise 4

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?

Of the previous two examples, exercise 9 had the smaller spread. If given the choice between a large or small spread I would choose the smaller spread as based on those examples it more closely reflects the total population and is less prone to outliers.