Data606 - Lab 5A

Cameron Smith

2020-10-02

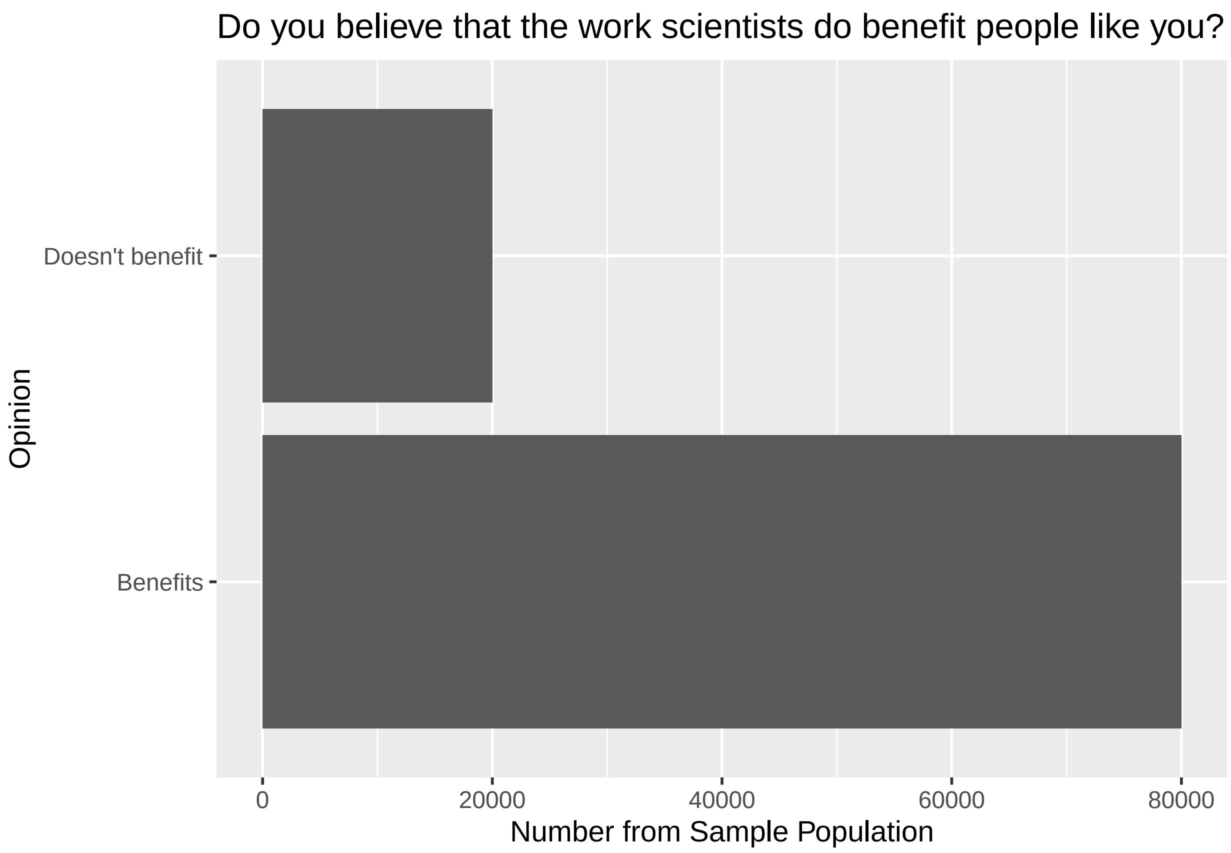
```
library(tidyverse)
library(openintro)
library(infer)
library(psych)
```

Exercise 1

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()
```



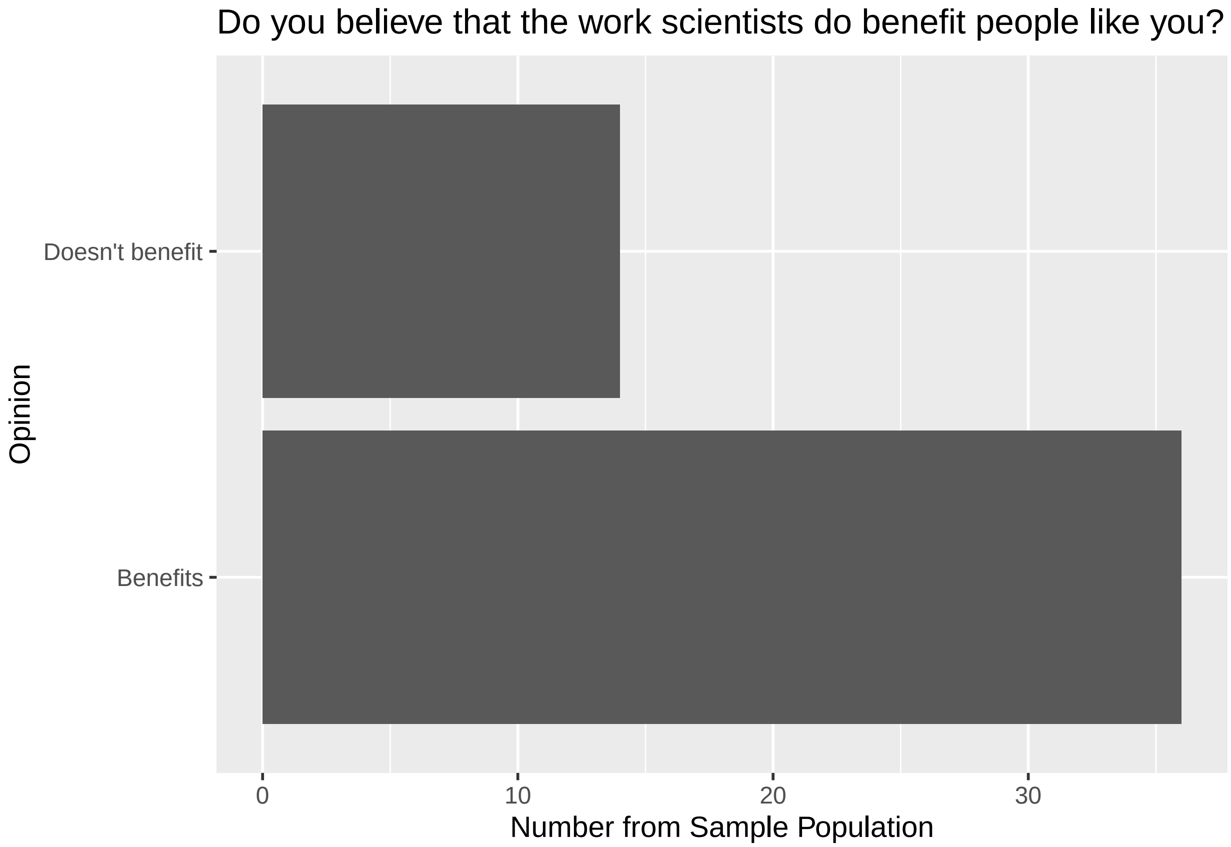
```r
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

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

```r
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()
```



Do you believe that the work scientists do benefit people like you?

```
# 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           36  0.72
## 2 Doesn’t benefit    14  0.28
```

Exercise 2

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           39  0.78
## 2 Doesn't benefit    11  0.22
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           40   0.8
## 2 Doesn't benefit    10   0.2
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits           38  0.76
## 2 Doesn't benefit    12  0.24
```

Exercise 3

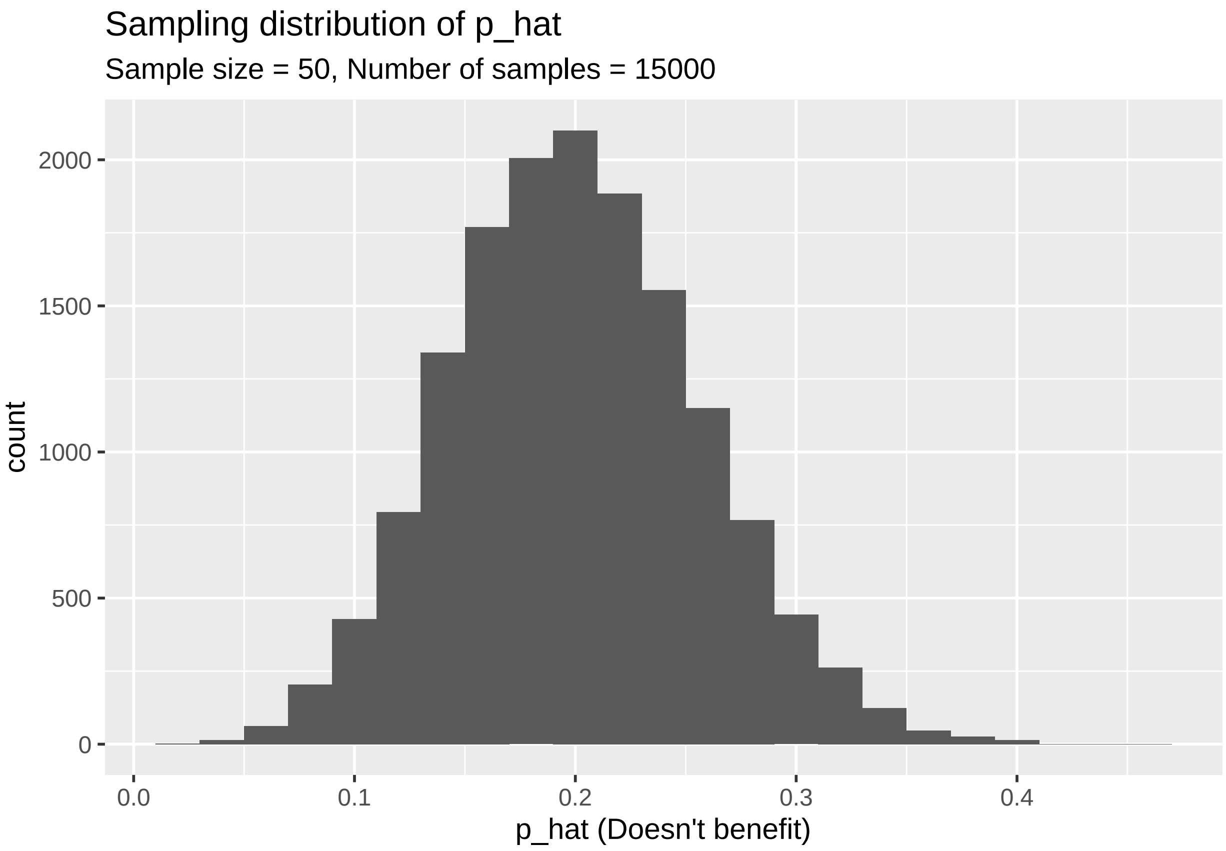
As illustrated above (in the code I used fr example 2), the sampe proportation will change each time unless the random seed is set. If we look at samples of 100 and 1000 then I would expect that 1000 would have a more accurate estimate of the population proportation. This is based on the law of large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

Exercise 4

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)
```

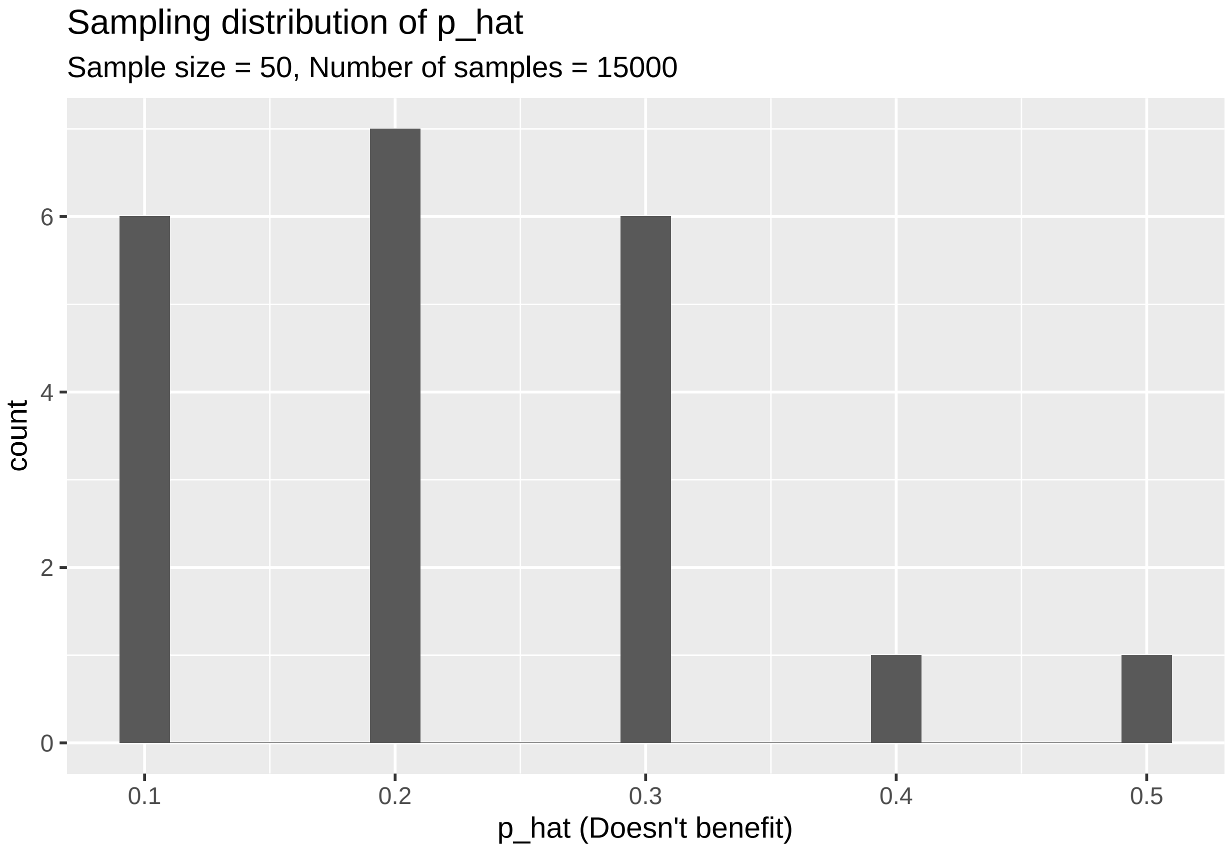
```
## [1] 15000
```

Exercise 5

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.

```r
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"
  )
```



Exercise 6

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 22 0.26 0.11   0.25    0.26 0.07 0.1 0.5   0.4 0.29    -0.84 0.02
```

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

```
##    vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 25 0.19 0.05   0.18    0.19 0.03 0.08 0.32  0.24 0.44    -0.05 0.01
```

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

```
##    vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 25  0.2 0.04    0.2     0.2 0.04 0.11 0.31   0.2 0.27      0.2 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 0.08 0.32  0.24 0.13    -0.29  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 0.07 0.37   0.3 0.15     0.02  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 0.04 0.38  0.34 0.17     0.02  0
```

Exercise 7

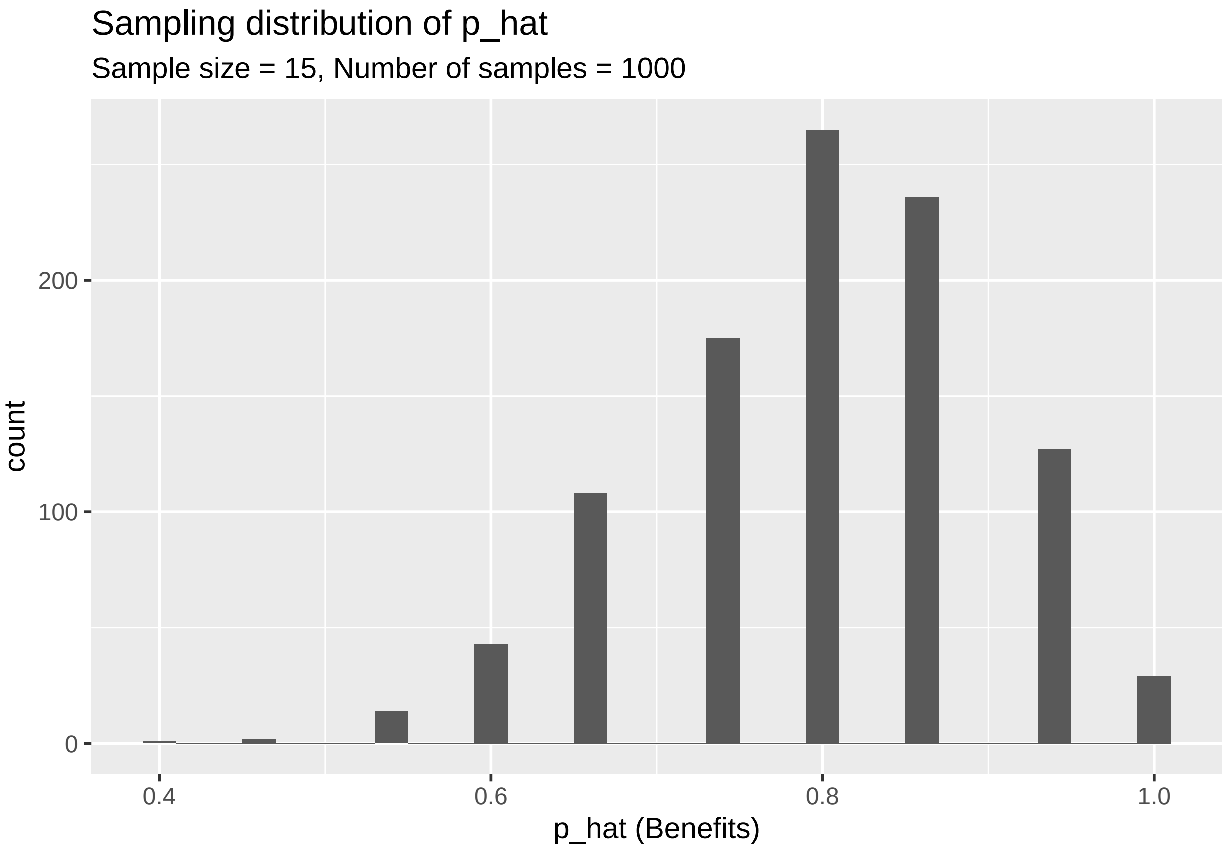
Based on the data and boxplot below we can see that the proportion of people who think the work that scientists do enhances their lives is 80% (this is the mean and median, and the standard error is 0).

```
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 0.4   1   0.6 -0.41        0  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"
  )
```



Exercise 8

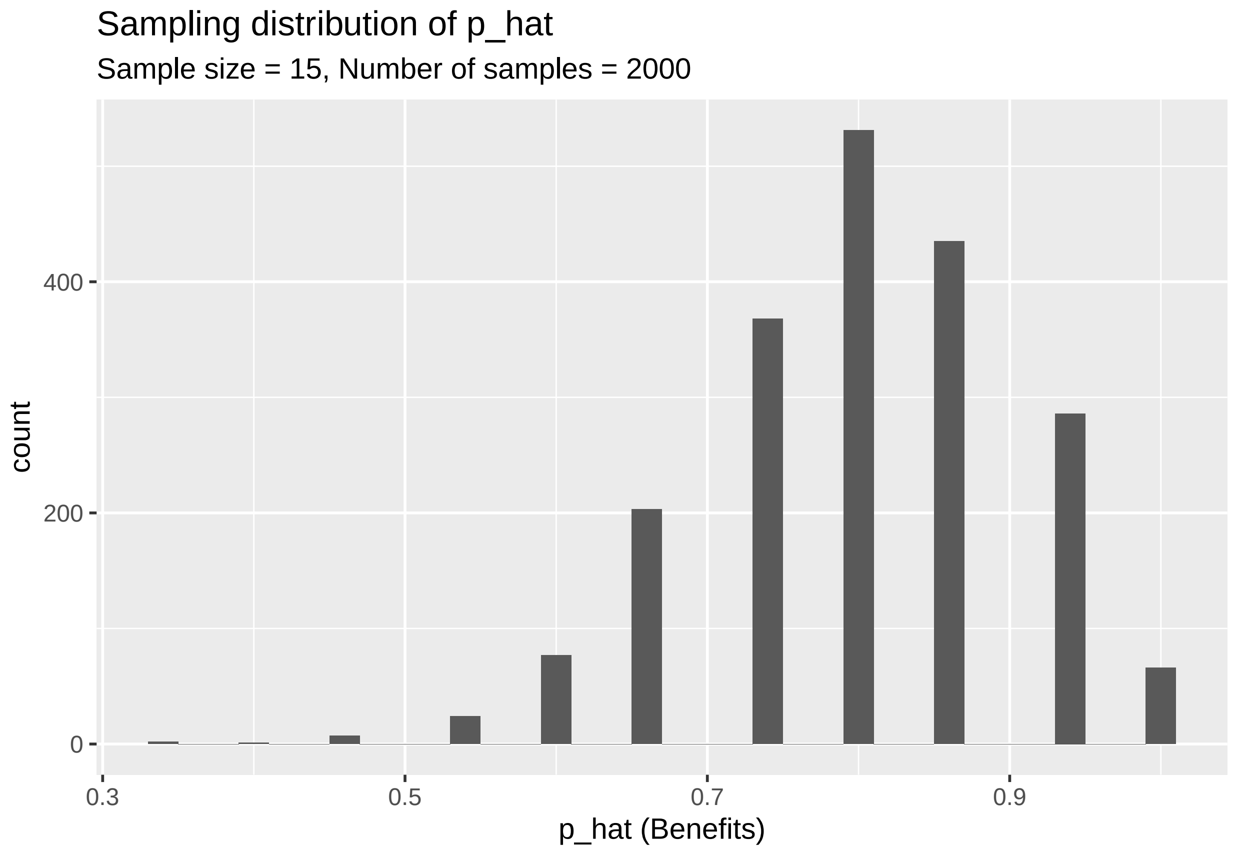
The distribution is normal, bimodal, and left-skewed. Based solely on looking at the histogram below, I would estimate the proportion of people who think the work that scientists do enhances their lives to be between 80% and 90%. The actual proportion is 80% as calculated using the describe() function 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 0.4   1   0.6 -0.41        0  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"
  )
```



Exercise 9

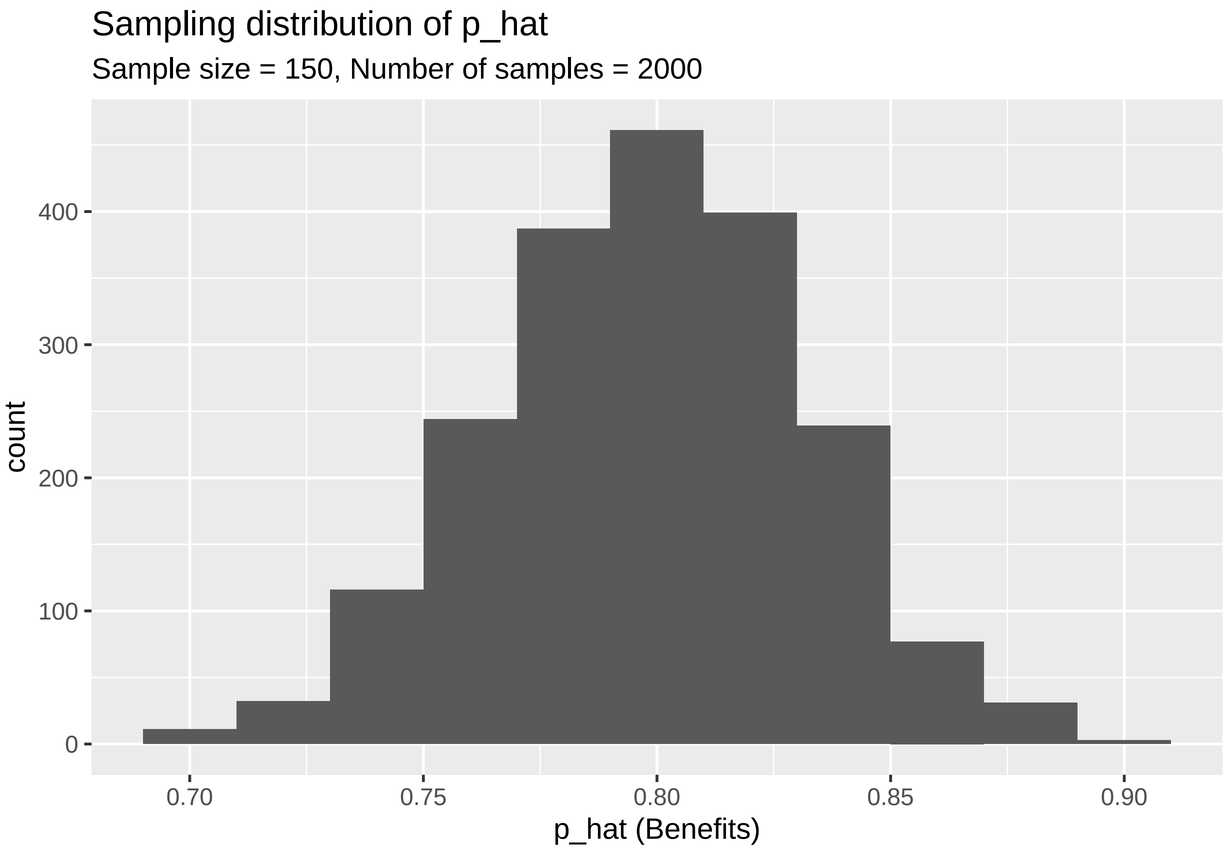
Based solely on looking at the histogram below, I would estimate the proportion of people who think the work that scientists do enhances their lives to be 80%. Unlike the previous example with a bimodal distribution, this is unimodal so the answer is pretty apparent.

```
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 0.69 0.89   0.2 -0.12    -0.13  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"
  )
```



Exercise 10

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.