Lab 5 (Part A)

Code

Julian Adames-Ng

2022-03-09

Hide

```
library(tidyverse)
library(openintro)
library(infer)
library(shiny)
```

Exercise 1

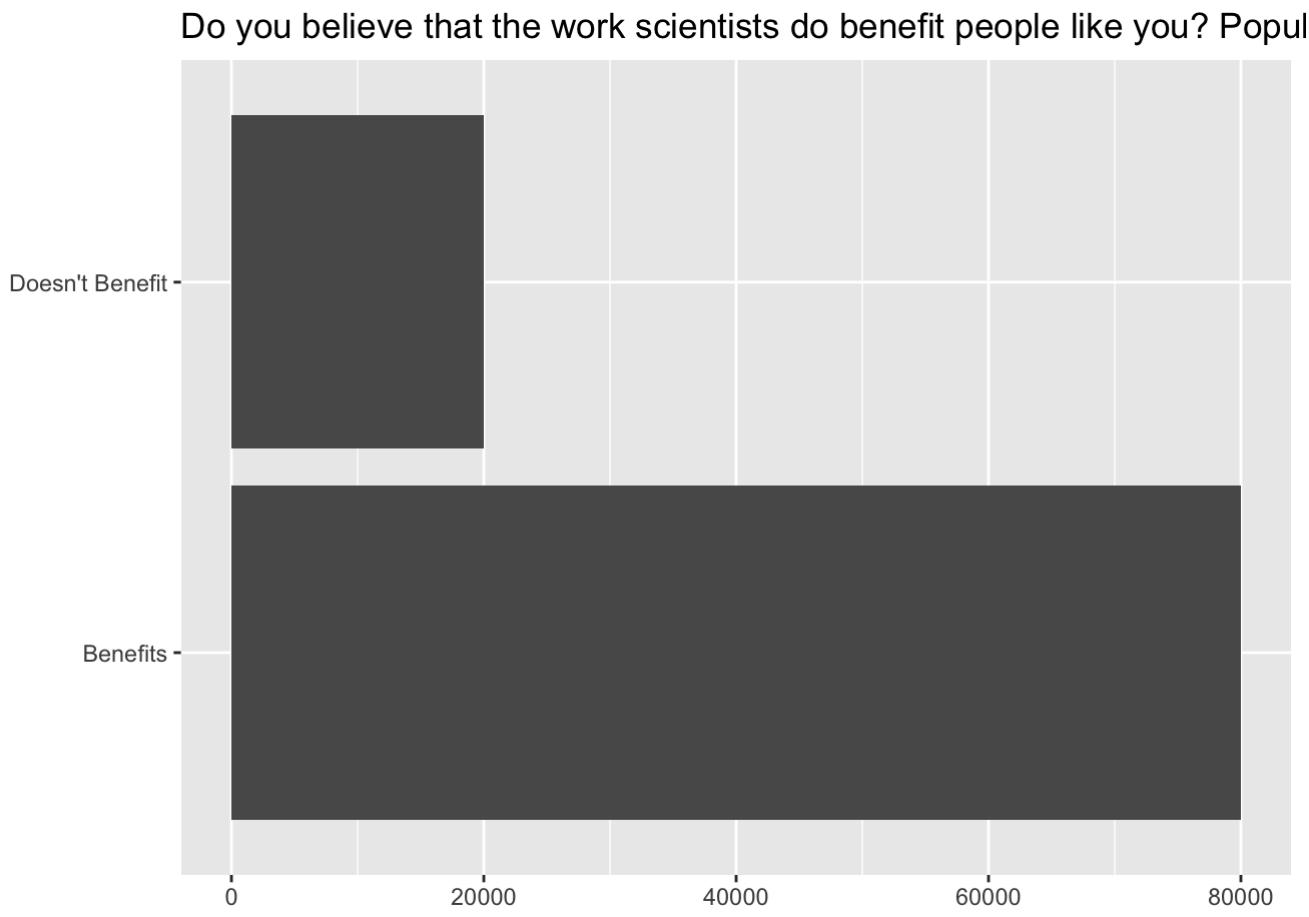
Describe the distribution of responses in this sample. How does it compare to the distribution of responses in the population. Hint: Although the sample_n function takes a random sample of observations (i.e. rows) from the dataset, you can still refer to the variables in the dataset with the same names. Code you presented earlier for visualizing and summarizing the population data will still be useful for the sample, however be careful to not label your proportion p since you're now calculating a sample statistic, not a population parameters. You can customize the label of the statistics to indicate that it comes from the sample.

-The responses in the sample yield 41 who believe that they benefit from scientists' work and 9 that do not believe so. This corresponds to a 0.82 and 0.18 as proportion values.

Hide

```
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? Po
        pulation"
  ) +
  coord_flip()
```



Hide

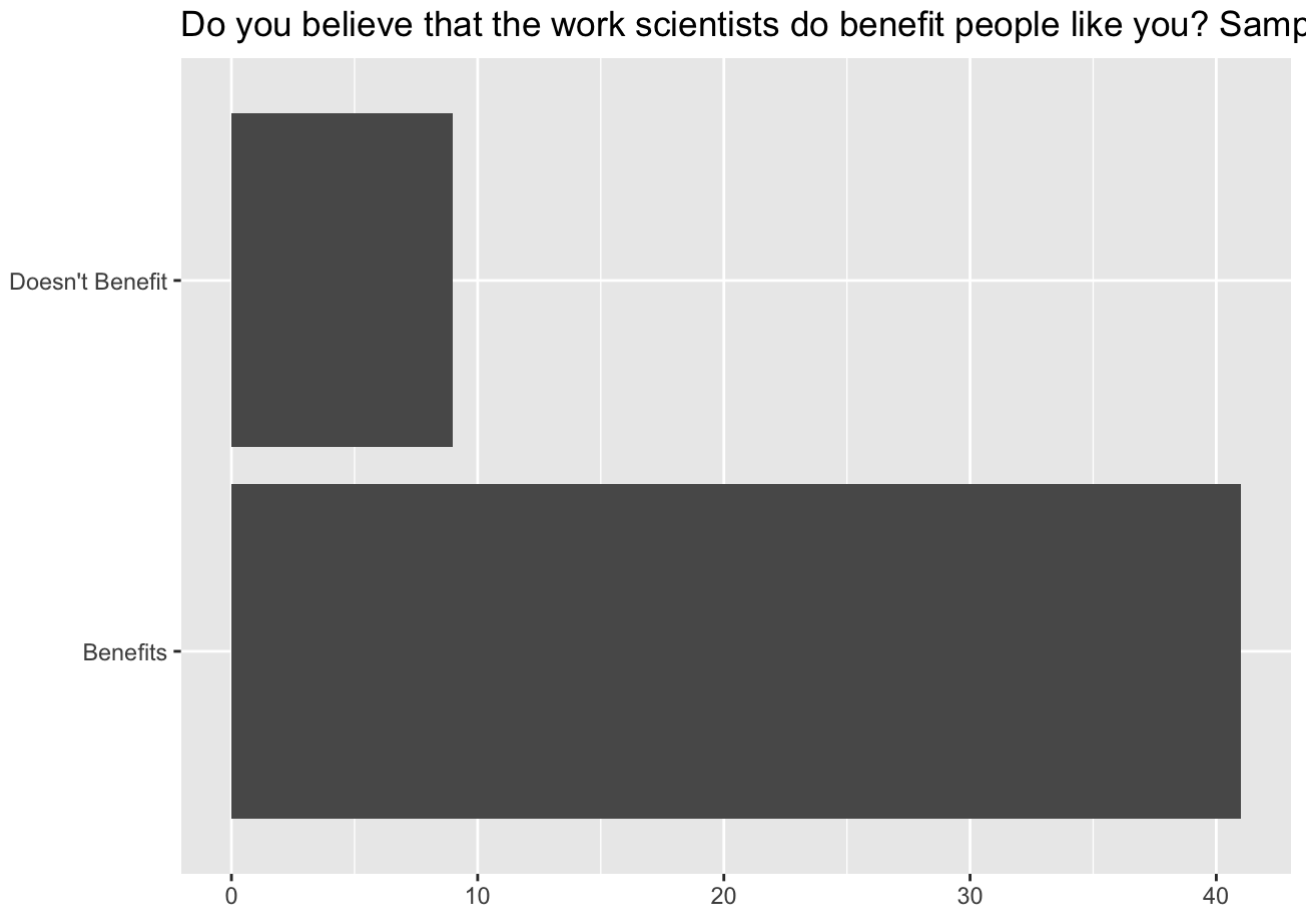
```
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n/sum(n))
```

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

Hide

```
#From sample1
set.seed(35797)
samp1 <- global_monitor %>%
  sample_n(50)

set.seed(35797)
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? Sa
        mple 1"
  ) +
  coord_flip()
```



Hide

```
set.seed(35797)
samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n/sum(n))
```

```
## # A tibble: 2 × 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           41  0.82
## 2 Doesn't Benefit     9  0.18
```

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 wouldn't expect the sample proportion to match the population proportion, but I would expect them to be somewhat similar.

...

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?

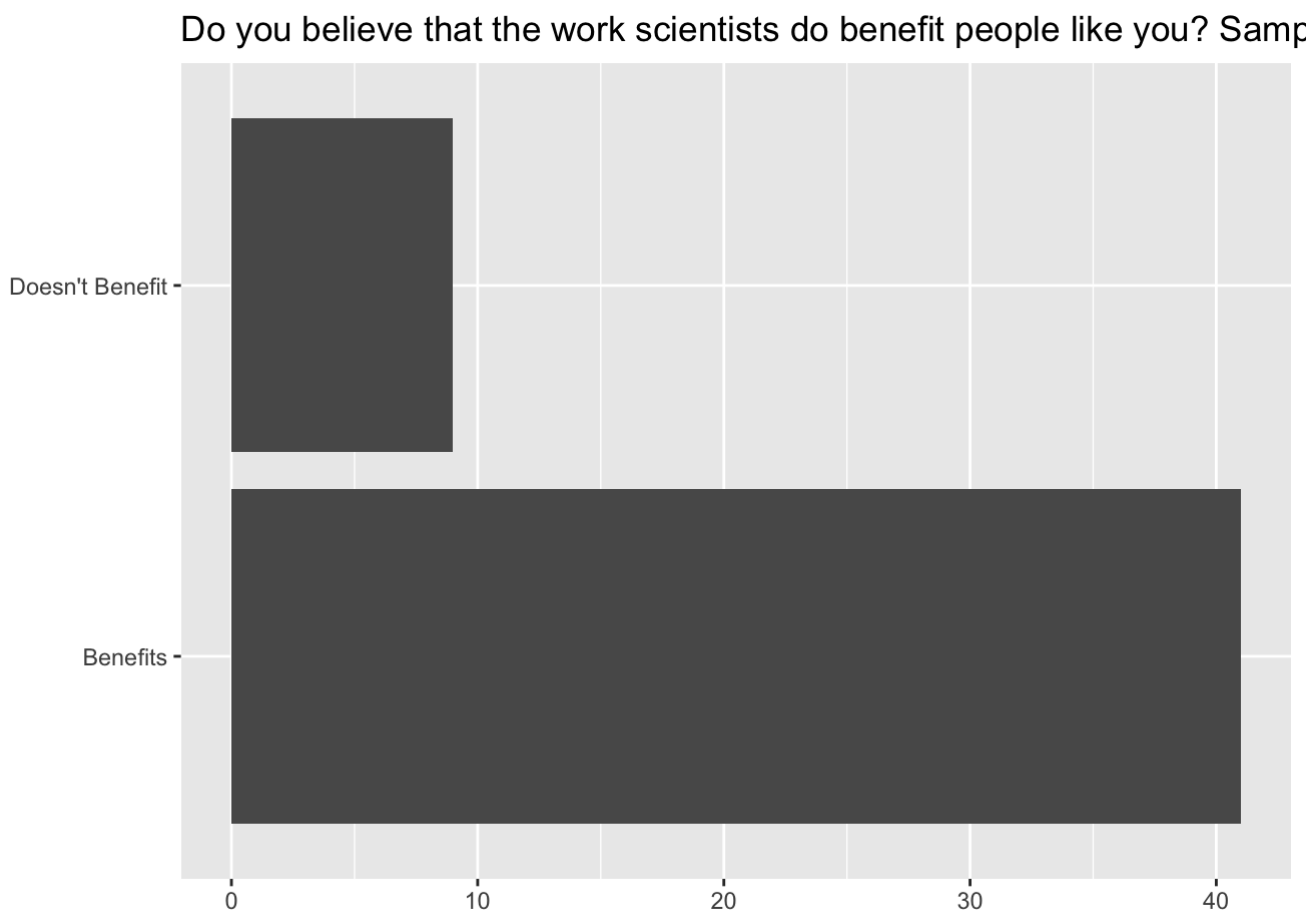
-The sample proportions do not match, but they are very similar in value. If we took samples with larger sample sizes, as the sizes get larger, the value of the sample proportion should approach the true value of the population proportion as per The Central Limit Theorem.

...

Hide

```
#From sample1
set.seed(35797)
samp1 <- global_monitor %>%
  sample_n(50)

set.seed(35797)
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? Sa
        mple 1"
  ) +
  coord_flip()
```



Hide

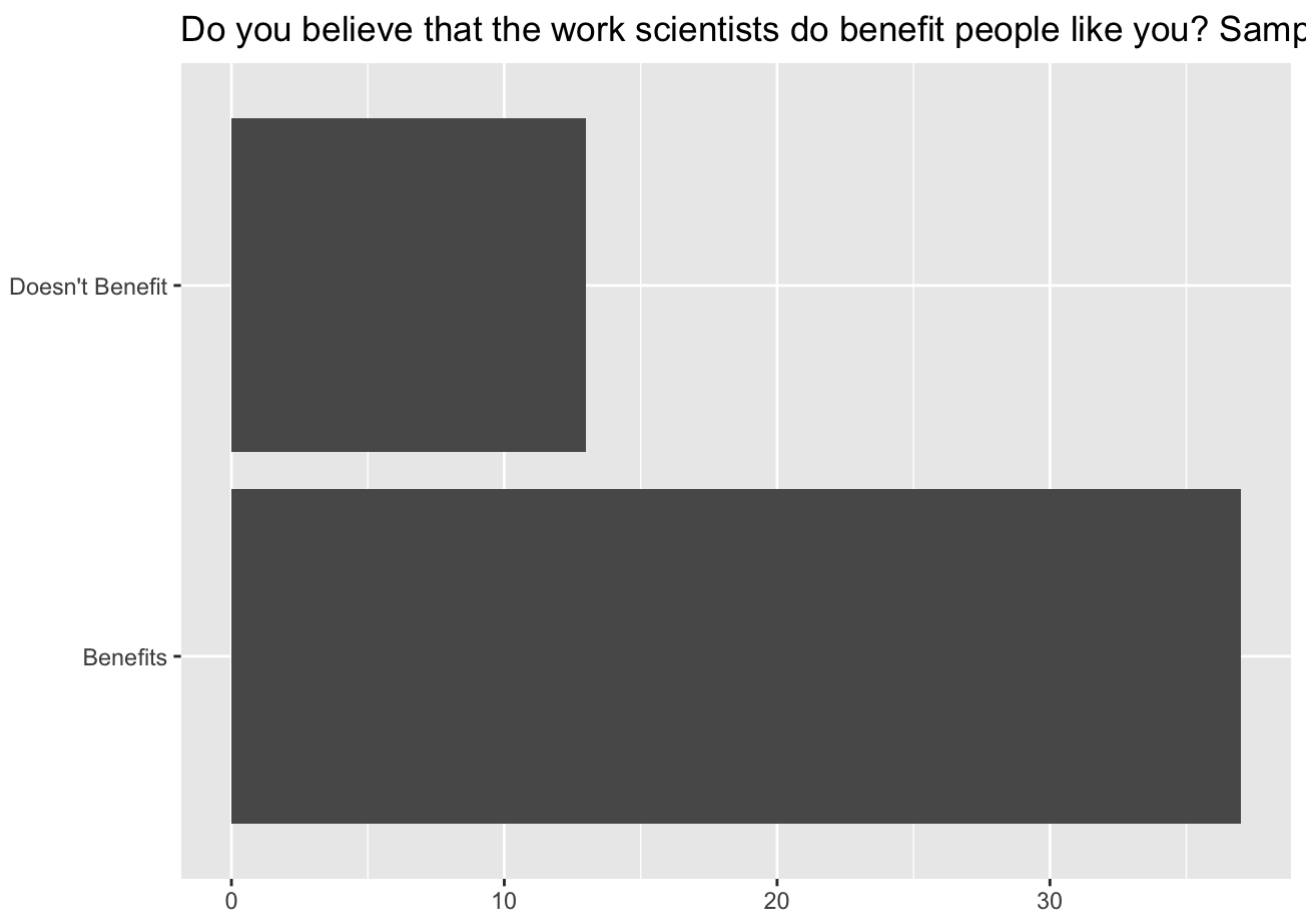
```
set.seed(35797)
samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n/sum(n))
```

```
## # A tibble: 2 × 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           41  0.82
## 2 Doesn't Benefit     9  0.18
```

Hide

```
#From sample2
set.seed(35798)
samp2 <- global_monitor %>%
  sample_n(50)

set.seed(35798)
ggplot(samp2, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? Sa
        mple 2"
  ) +
  coord_flip()
```



Hide

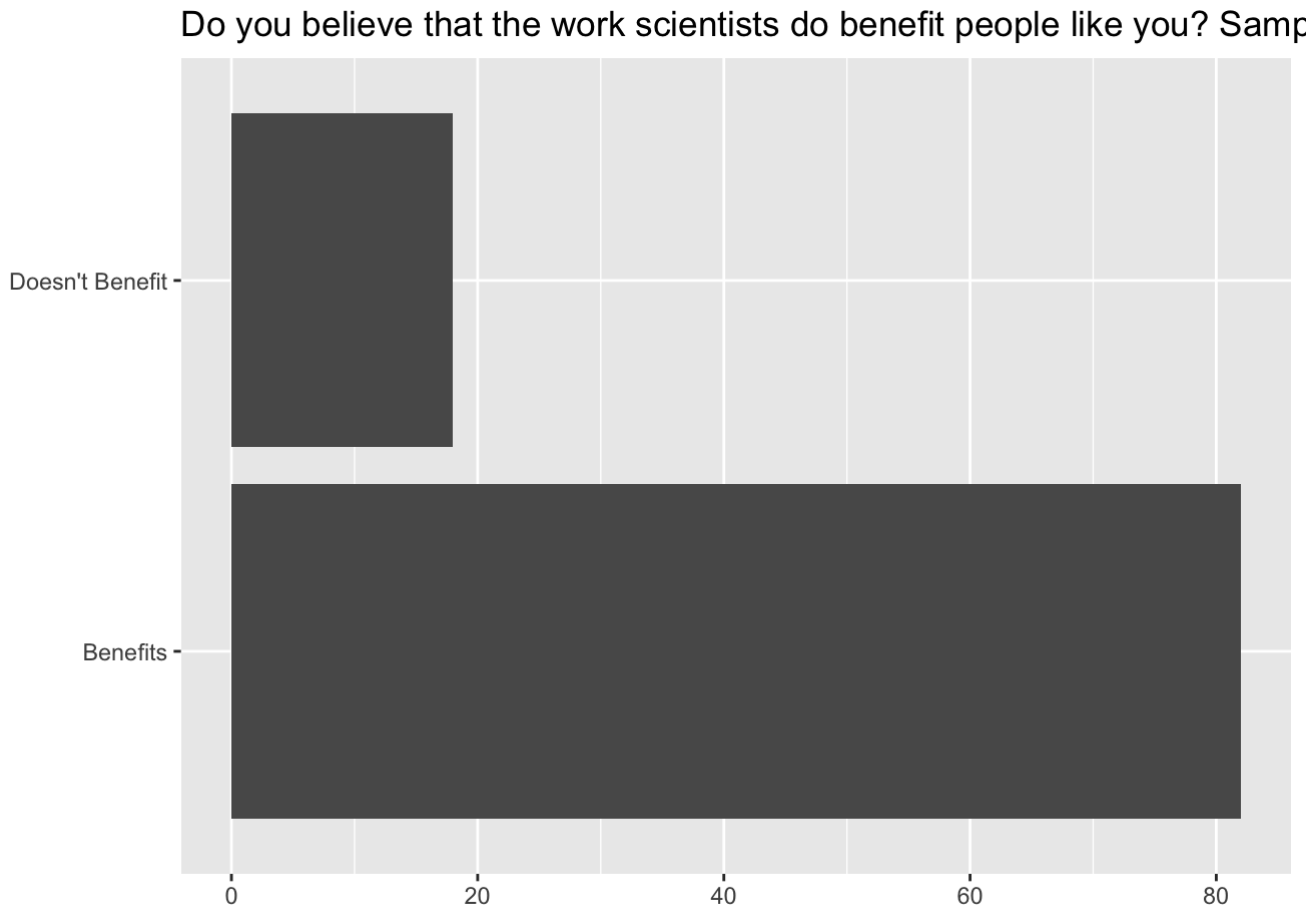
```
set.seed(35798)
samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n/sum(n))
```

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

Hide

```
#From sample3
set.seed(35799)
samp3 <- global_monitor %>%
  sample_n(100)

set.seed(35799)
ggplot(samp3, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? Sa
        mple 3"
  ) +
  coord_flip()
```



Hide

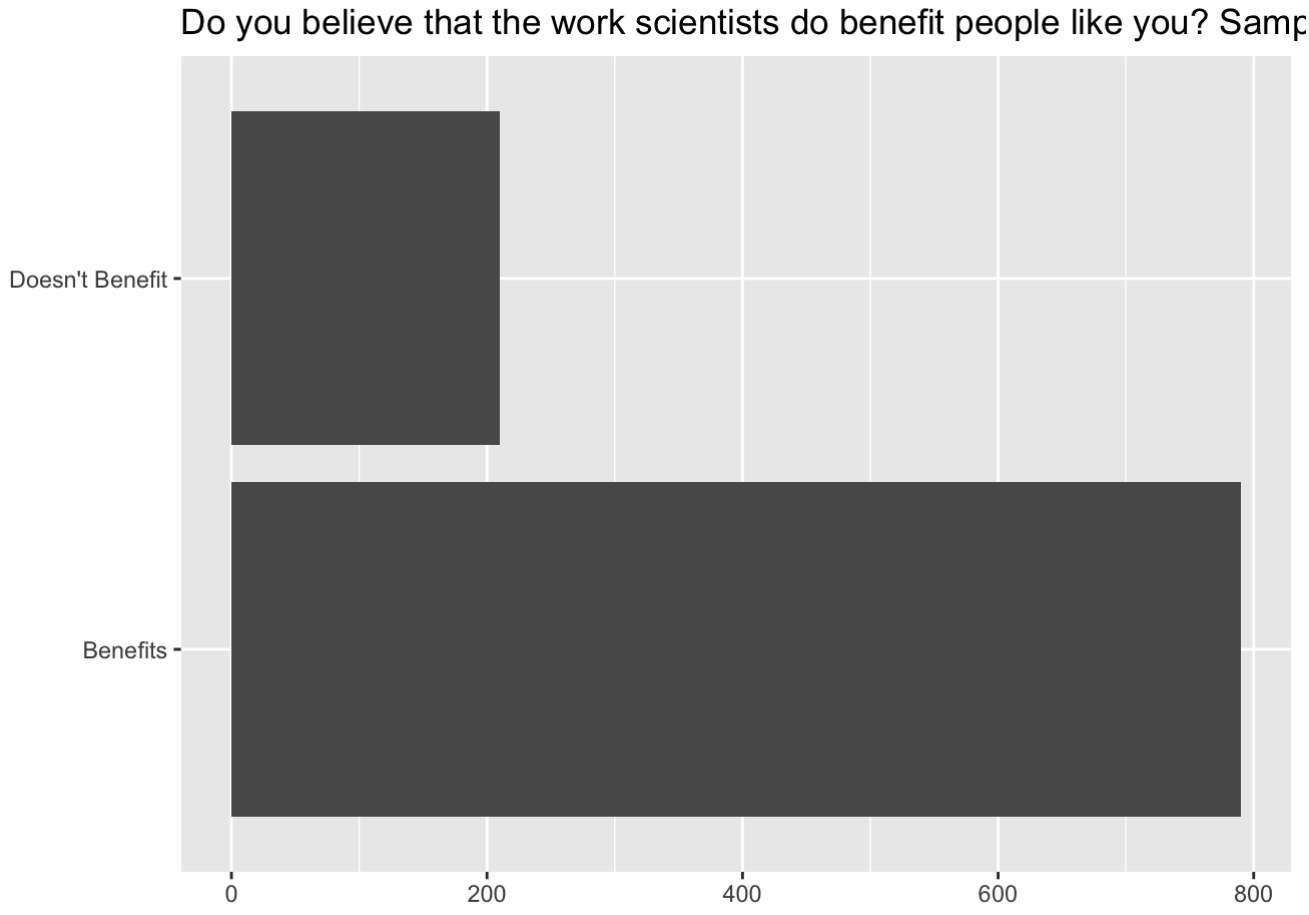
```
set.seed(35799)
samp3 %>%
  count(scientist_work) %>%
  mutate(p_hat = n/sum(n))
```

```
## # A tibble: 2 × 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           82  0.82
## 2 Doesn't Benefit    18  0.18
```

Hide

```
#From sample4
set.seed(35790)
samp4 <- global_monitor %>%
  sample_n(1000)

set.seed(35790)
ggplot(samp4, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? Sa
        mple 4"
  ) +
  coord_flip()
```



Hide

```
set.seed(35790)
samp4 %>%
  count(scientist_work) %>%
  mutate(p_hat = n/sum(n))
```

```
## # A tibble: 2 × 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          790  0.79
## 2 Doesn't Benefit   210  0.21
```

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 15000 elements in sample_props50. The sampling distribution looks normal and has a center around 0.2. The calculated mean is 0.200392 which is a close estimate.

...

Hide

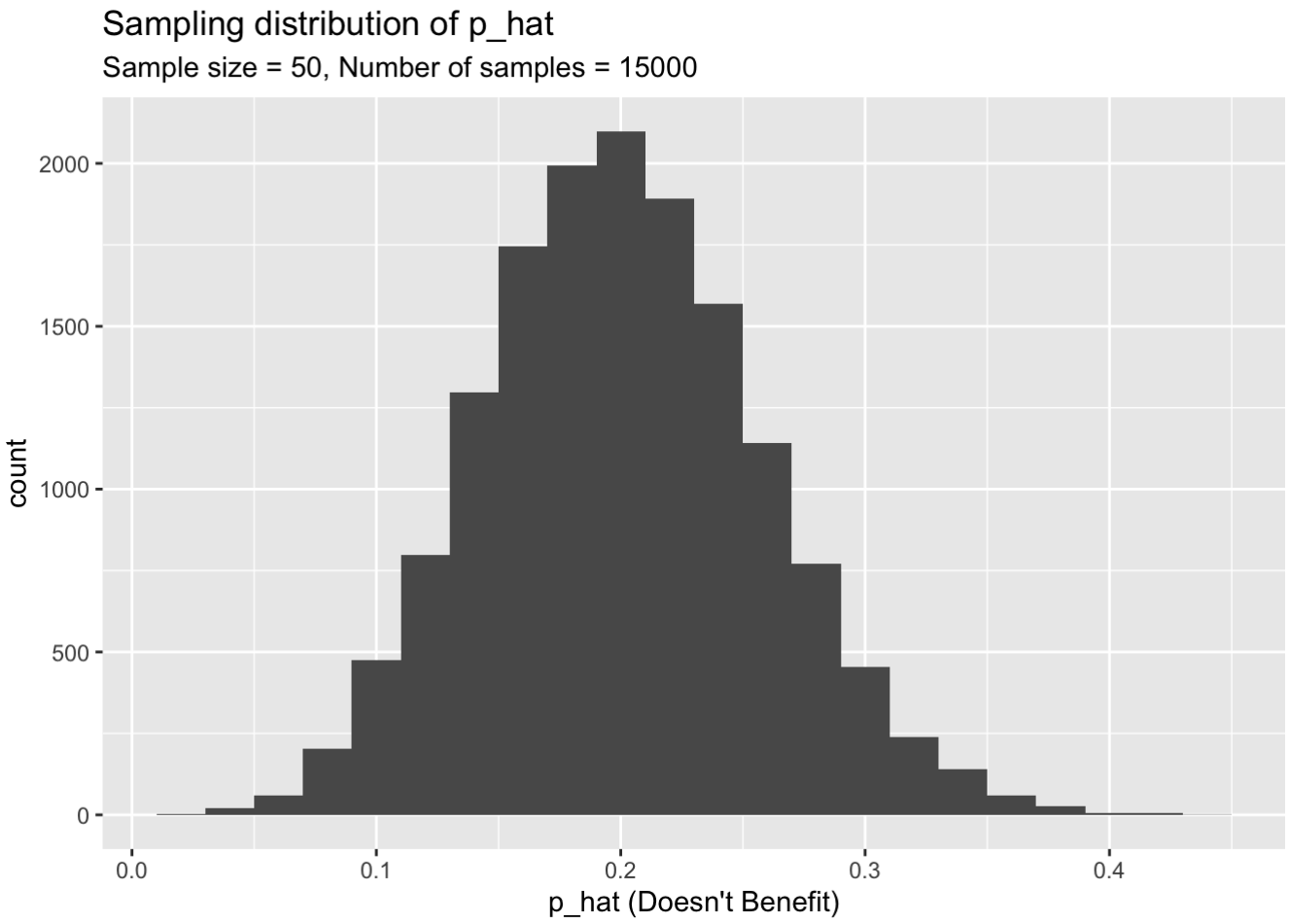
```
set.seed(8675)
sample_props50 <- global_monitor %>%
                  rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%
                  count(scientist_work) %>%
                  mutate(p_hat = n/sum(n)) %>%
                  filter(scientist_work == "Doesn't Benefit")

set.seed(8675)
sample_props50
```

```
## # A tibble: 15,000 × 4
## # Groups:   replicate [15,000]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't Benefit     5  0.1
##  2         2 Doesn't Benefit    10  0.2
##  3         3 Doesn't Benefit     7  0.14
##  4         4 Doesn't Benefit     8  0.16
##  5         5 Doesn't Benefit    10  0.2
##  6         6 Doesn't Benefit    10  0.2
##  7         7 Doesn't Benefit    16  0.32
##  8         8 Doesn't Benefit    15  0.3
##  9         9 Doesn't Benefit    12  0.24
## 10        10 Doesn't Benefit    11  0.22
## # … with 14,990 more rows
```

Hide

```
set.seed(8675)
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"
  )
```



Hide

```
#number of elements corresponds to row
nrow(sample_props50)
```

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

Hide

```
#mean of sampling distribution
set.seed(8675)
sample_props50 %>%
  summarise(mean_sp   = mean(sample_props50$p_hat),
            n         = n())
```

```
## # A tibble: 15,000 × 3
##    replicate mean_sp     n
##        <int>   <dbl> <int>
##  1         1   0.200     1
##  2         2   0.200     1
##  3         3   0.200     1
##  4         4   0.200     1
##  5         5   0.200     1
##  6         6   0.200     1
##  7         7   0.200     1
##  8         8   0.200     1
##  9         9   0.200     1
## 10        10   0.200     1
## # … with 14,990 more rows
```

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?

-There are 25 observations in sample_props_small.

...

Hide

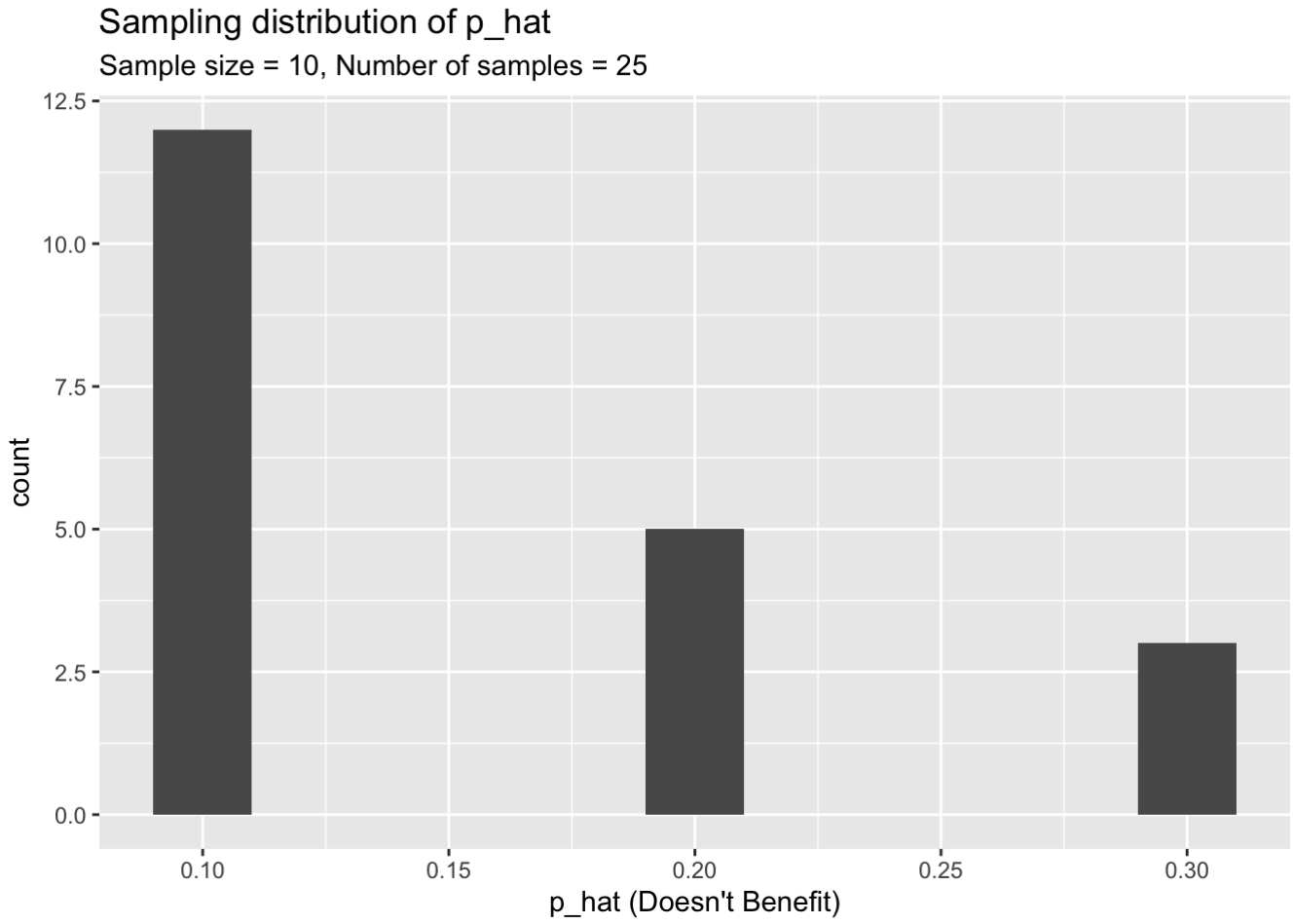
```
set.seed(8676)
sample_props_small <- global_monitor %>%
                  rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
                  count(scientist_work) %>%
                  mutate(p_hat = n/sum(n)) %>%
                  filter(scientist_work == "Doesn't Benefit")

set.seed(8676)
sample_props_small
```

```
## # A tibble: 20 × 4
## # Groups:   replicate [20]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't Benefit     1   0.1
##  2         2 Doesn't Benefit     1   0.1
##  3         3 Doesn't Benefit     1   0.1
##  4         4 Doesn't Benefit     1   0.1
##  5         6 Doesn't Benefit     1   0.1
##  6         8 Doesn't Benefit     2   0.2
##  7        10 Doesn't Benefit     3   0.3
##  8        11 Doesn't Benefit     2   0.2
##  9        12 Doesn't Benefit     1   0.1
## 10        13 Doesn't Benefit     1   0.1
## 11        14 Doesn't Benefit     2   0.2
## 12        15 Doesn't Benefit     1   0.1
## 13        17 Doesn't Benefit     1   0.1
## 14        18 Doesn't Benefit     1   0.1
## 15        19 Doesn't Benefit     3   0.3
## 16        20 Doesn't Benefit     3   0.3
## 17        22 Doesn't Benefit     1   0.1
## 18        23 Doesn't Benefit     2   0.2
## 19        24 Doesn't Benefit     1   0.1
## 20        25 Doesn't Benefit     2   0.2
```

Hide

```
set.seed(8676)
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 = 10, Number of samples = 25"
  )
```



Hide

```
#number of elements corresponds to row
nrow(sample_props_small)
```

```
## [1] 20
```

Hide

```
#mean of sampling distribution
set.seed(8676)
sample_props_small %>%
  summarise(mean_sps   = mean(sample_props_small$p_hat),
            n          = n())
```

```
## # A tibble: 20 × 3
##    replicate mean_sps     n
##        <int>    <dbl> <int>
##  1         1    0.155     1
##  2         2    0.155     1
##  3         3    0.155     1
##  4         4    0.155     1
##  5         6    0.155     1
##  6         8    0.155     1
##  7        10    0.155     1
##  8        11    0.155     1
##  9        12    0.155     1
## 10        13    0.155     1
## 11        14    0.155     1
## 12        15    0.155     1
## 13        17    0.155     1
## 14        18    0.155     1
## 15        19    0.155     1
## 16        20    0.155     1
## 17        22    0.155     1
## 18        23    0.155     1
## 19        24    0.155     1
## 20        25    0.155     1
```

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.)

-Each observation represents an individual sample. As we increase the sample size, the distribution seems to change from right skewed to normal and the mean seems to approach 0.2 or so.

...

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?

-The best point estimate for the population proportion is about 0.78.

...

Hide

```
set.seed(27)

samp15 <- global_monitor %>%
  sample_n(27)

samp15 %>%
  count(scientist_work) %>%
  mutate(p_hat15 = n /sum(n))
```

```
## # A tibble: 2 × 3
##   scientist_work      n p_hat15
##   <chr>           <int>   <dbl>
## 1 Benefits           21   0.778
## 2 Doesn't Benefit     6   0.222
```

Exercise 8

Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enhances 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.

-The point estimate is 0.7984 which is close to the true proportion value of 0.8.

...

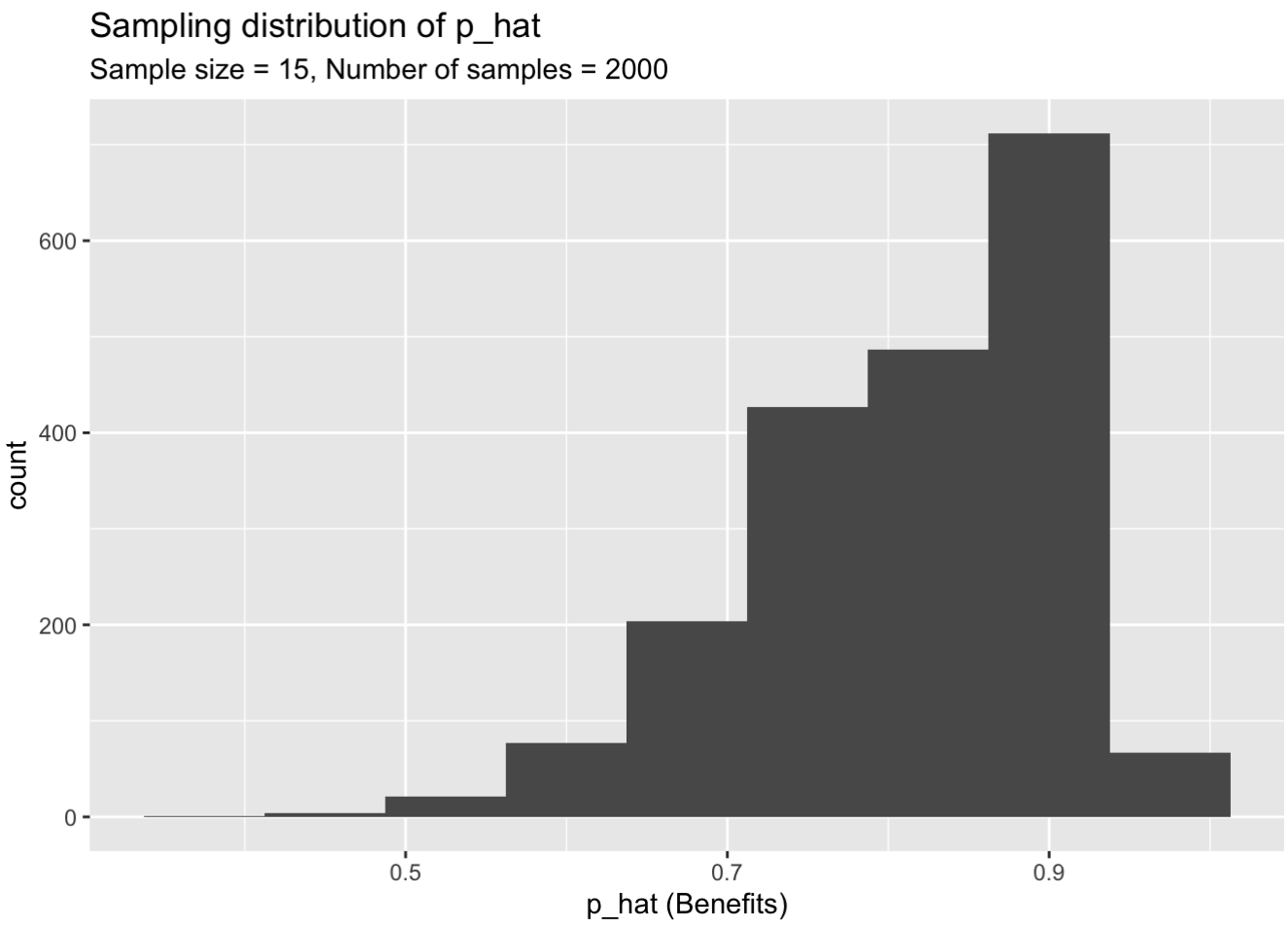
Hide

```
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")
sample_props15
```

```
## # A tibble: 2,000 × 4
## # Groups:   replicate [2,000]
##    replicate scientist_work     n p_hat
##        <int> <chr>          <int> <dbl>
##  1         1 Benefits          11 0.733
##  2         2 Benefits          15 1
##  3         3 Benefits          13 0.867
##  4         4 Benefits           8 0.533
##  5         5 Benefits          13 0.867
##  6         6 Benefits          12 0.8
##  7         7 Benefits          12 0.8
##  8         8 Benefits          12 0.8
##  9         9 Benefits          13 0.867
## 10        10 Benefits          13 0.867
## # … with 1,990 more rows
```

Hide

```
ggplot(data = sample_props15, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.075) +
  labs(x = "p_hat (Benefits)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 15, Number of samples = 2000")
```



Hide

```
mean(sample_props15$p_hat)
```

```
## [1] 0.7997667
```

Hide

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

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

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?

-The shape of the distribution looks more normal than anything. I'd guess that the true proportion is somewhere near 0.8 based on the histogram plot.

...

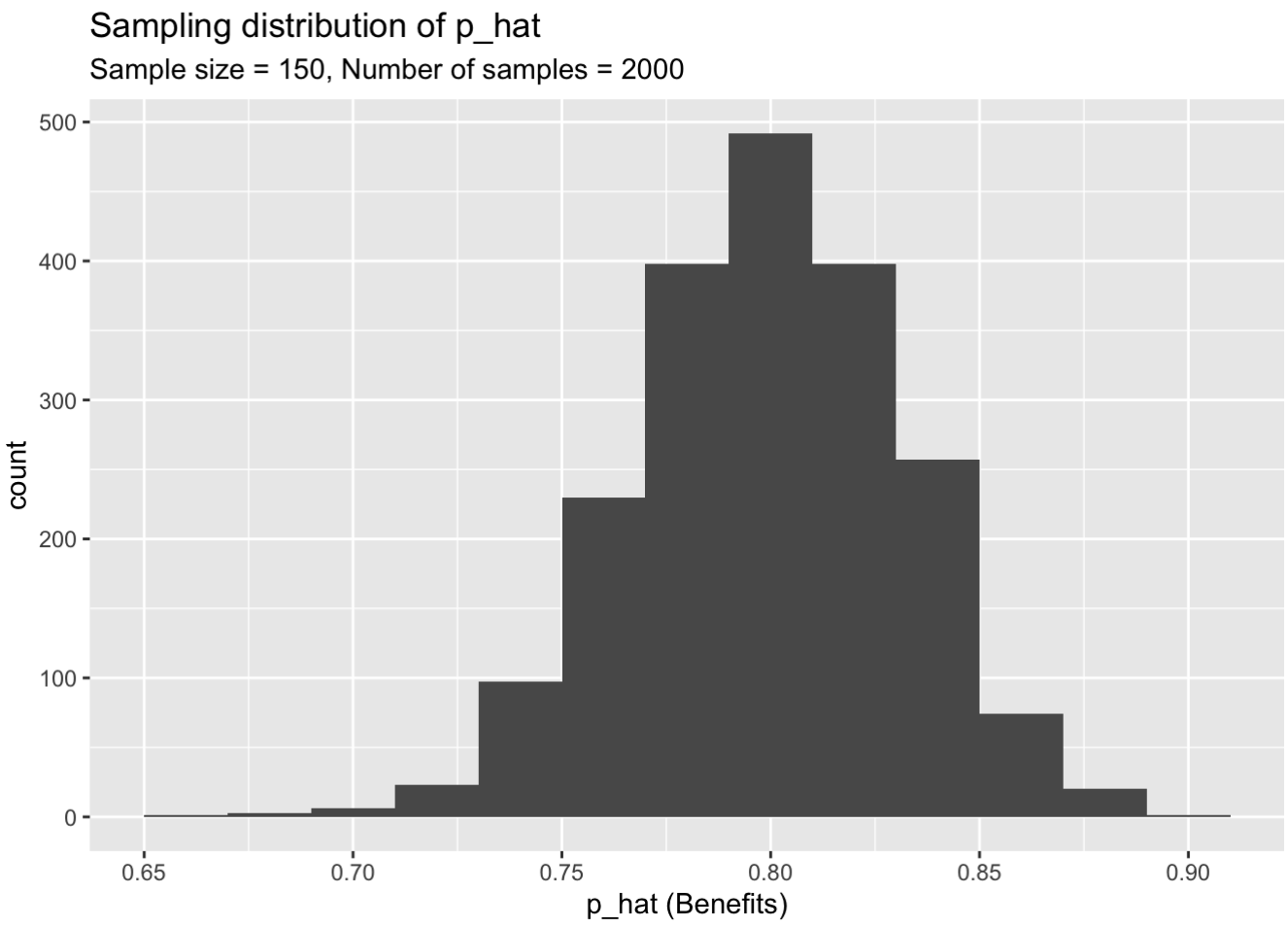
Hide

```
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")
sample_props150
```

```
## # A tibble: 2,000 × 4
## # Groups:   replicate [2,000]
##    replicate scientist_work     n p_hat
##        <int> <chr>          <int> <dbl>
##  1         1 Benefits         110 0.733
##  2         2 Benefits         121 0.807
##  3         3 Benefits         121 0.807
##  4         4 Benefits         124 0.827
##  5         5 Benefits         123 0.82
##  6         6 Benefits         119 0.793
##  7         7 Benefits         125 0.833
##  8         8 Benefits         119 0.793
##  9         9 Benefits         123 0.82
## 10        10 Benefits         119 0.793
## # … with 1,990 more rows
```

Hide

```
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")
```



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 distribution with the larger sampling size has a smaller spread. In this example, I'd prefer a smaller spread since it involves larger samples.

...