Data 606 Lab 5 - Foundations for statistical inference - Sampling distributions

Enid Roman

2022-10-06

```
set.seed(500)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

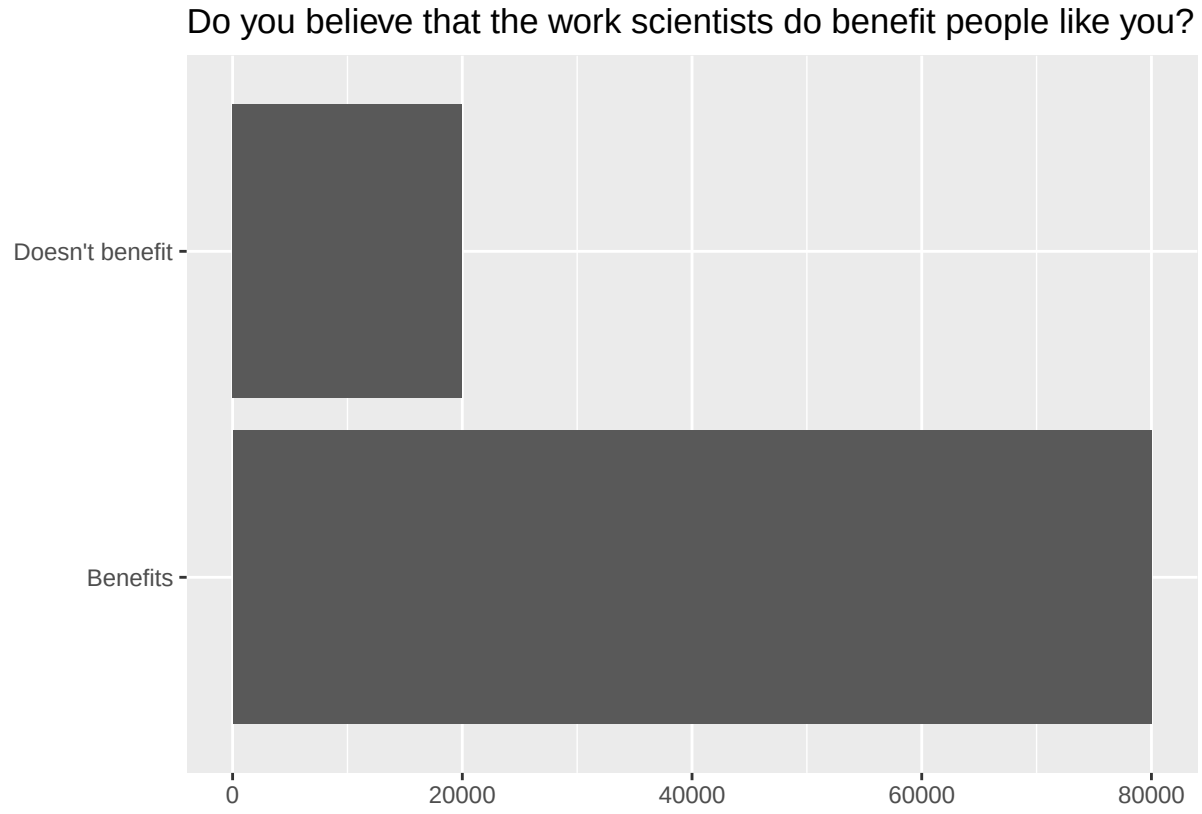
```
library(openintro)
```

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

```
library(infer)
```

```
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
```

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

Exercise 1

```
samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

The distribution of responses in this sample compare to the distribution of responses in the population sample is even though Benefits is a little lower in the sample compare to the population and Doesn't benefit is a little higher in the sample compare to the population both data sets shows that close to 20% of people believe that the work of scientists is not beneficial.

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           43  0.86
## 2 Doesn't benefit     7  0.14
```

Exercise 2

I would not expect the sample proportion to match the sample proportion of another student's sample because the sample is randomly selected so the sample result would always come out a little different everytime you run the sample code. I would expect the proportion to be not much of a difference. Maybe about 5% to 10% difference.

Exercise 3

```
samp2 <- global_monitor %>%
  sample_n(50)


samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

The sample proportion of samp2 is a little different compare with that of samp1. The proportion of Benefits has increased in samp2. Below you can see that the more sample you add, 100 and 1000, the more accurate is the measurement. So adding 1000 samples will give a more accurate measurement.

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           42  0.84
## 2 Doesn't benefit     8  0.16
```

```
samp3 <- global_monitor %>%
  sample_n(100)


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           78  0.78
## 2 Doesn't benefit    22  0.22
```

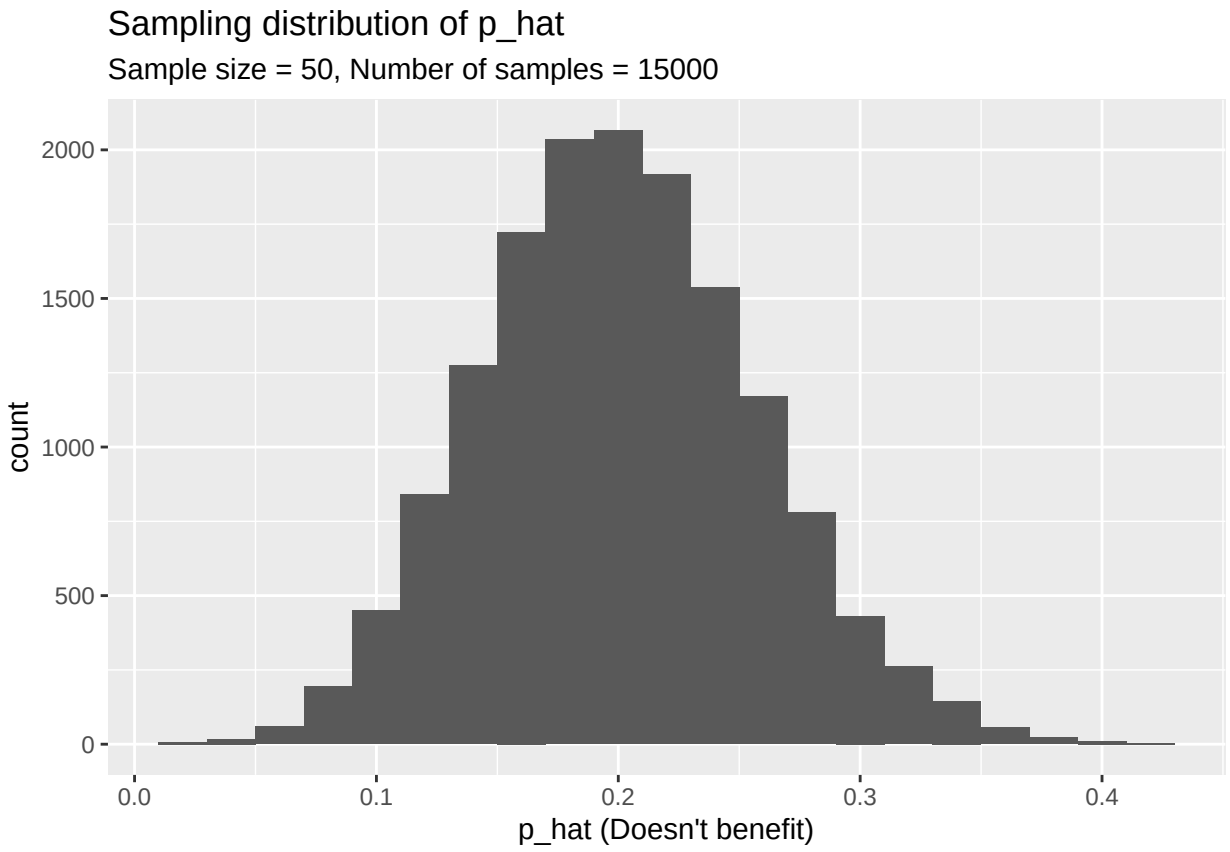
```r
samp4 <- global_monitor %>%
  sample_n(1000)


samp4 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          780  0.78
## 2 Doesn't benefit   220  0.22
```

```r
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")
```

```r
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"
  )
```



Exercise 4

```
global_monitor %>%
  sample_n(size = 50, replace = TRUE) %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Doesn't benefit")
```

There are approx. a little over 15000 elements in sample_props50. The sampling distribution looks like a normal distribution with no skewed. The center is .2 mean.

```
## # A tibble: 1 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Doesn't benefit    14  0.28
```

Exercise 5

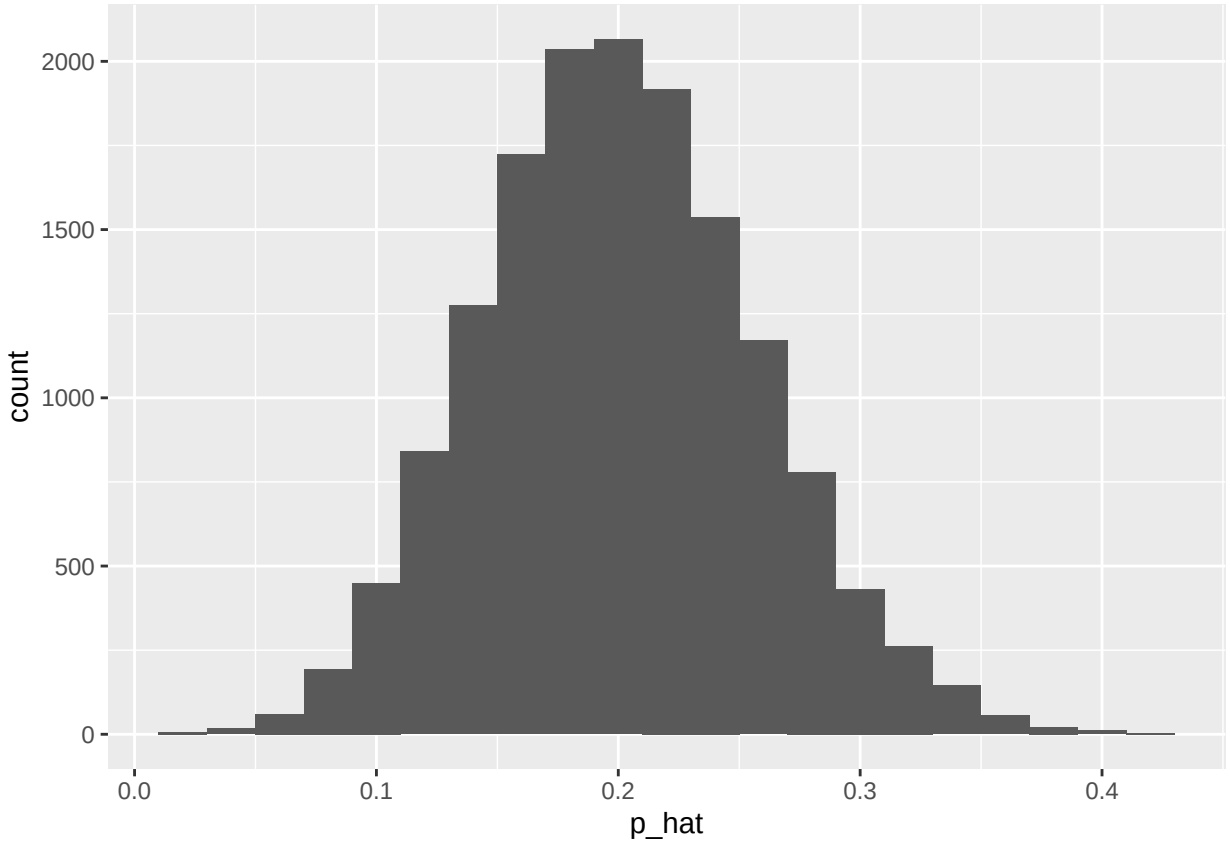
```
sample_props_small <- global_monitor %>%
                    rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
```

```
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
sample_props_small
```

There are 25 observations in this object called sample_props_small. The observation represent the sample propotion (n) which makes the sample distribution (p_hat)

```
## # A tibble: 25 x 4
## # Groups:   replicate [25]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     2   0.2
##  2         2 Doesn't benefit     3   0.3
##  3         3 Doesn't benefit     3   0.3
##  4         4 Doesn't benefit     1   0.1
##  5         5 Doesn't benefit     1   0.1
##  6         6 Doesn't benefit     2   0.2
##  7         7 Doesn't benefit     1   0.1
##  8         8 Doesn't benefit     2   0.2
##  9         9 Doesn't benefit     2   0.2
## 10        10 Doesn't benefit     1   0.1
## # ... with 15 more rows
```

```
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02)
```

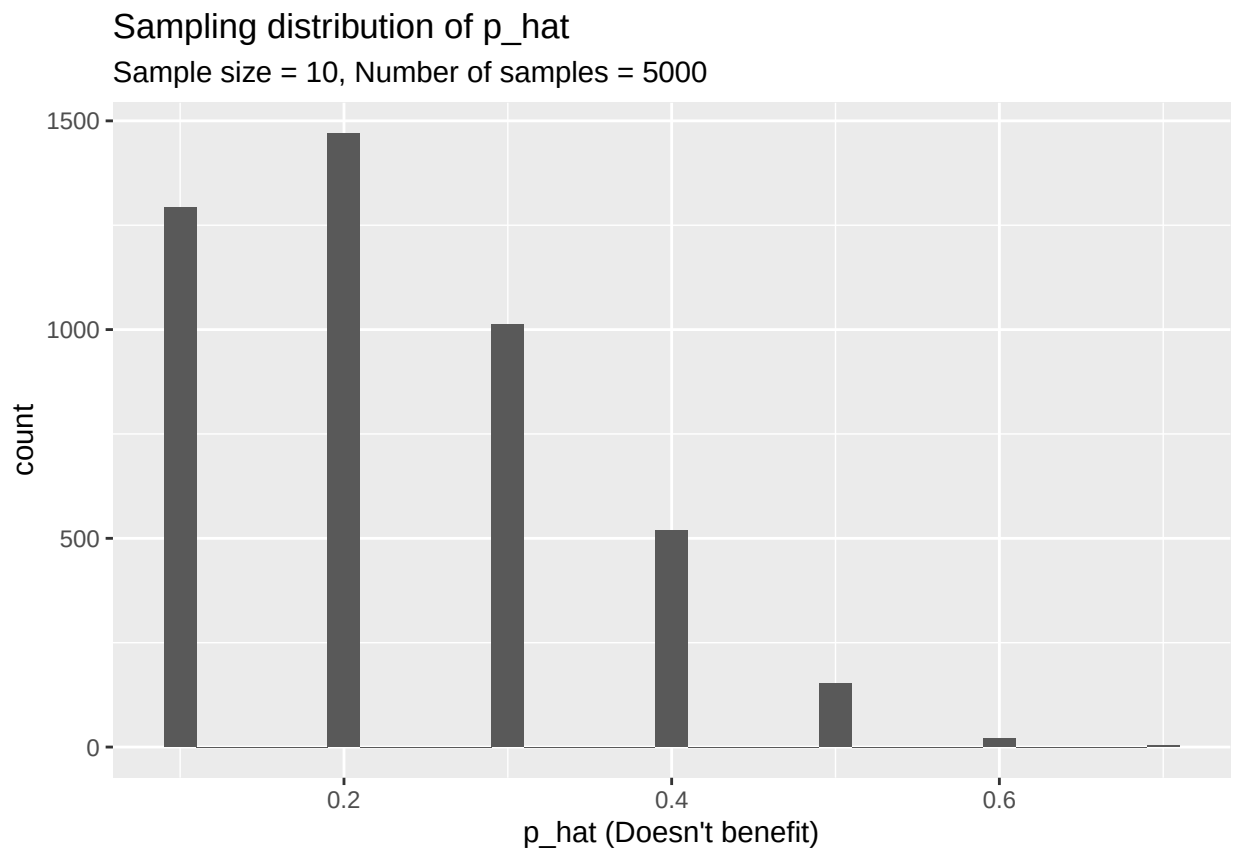


Exercise 6

```r
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")

ggplot(data = sample_props10, 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 = 5000"
  )
```

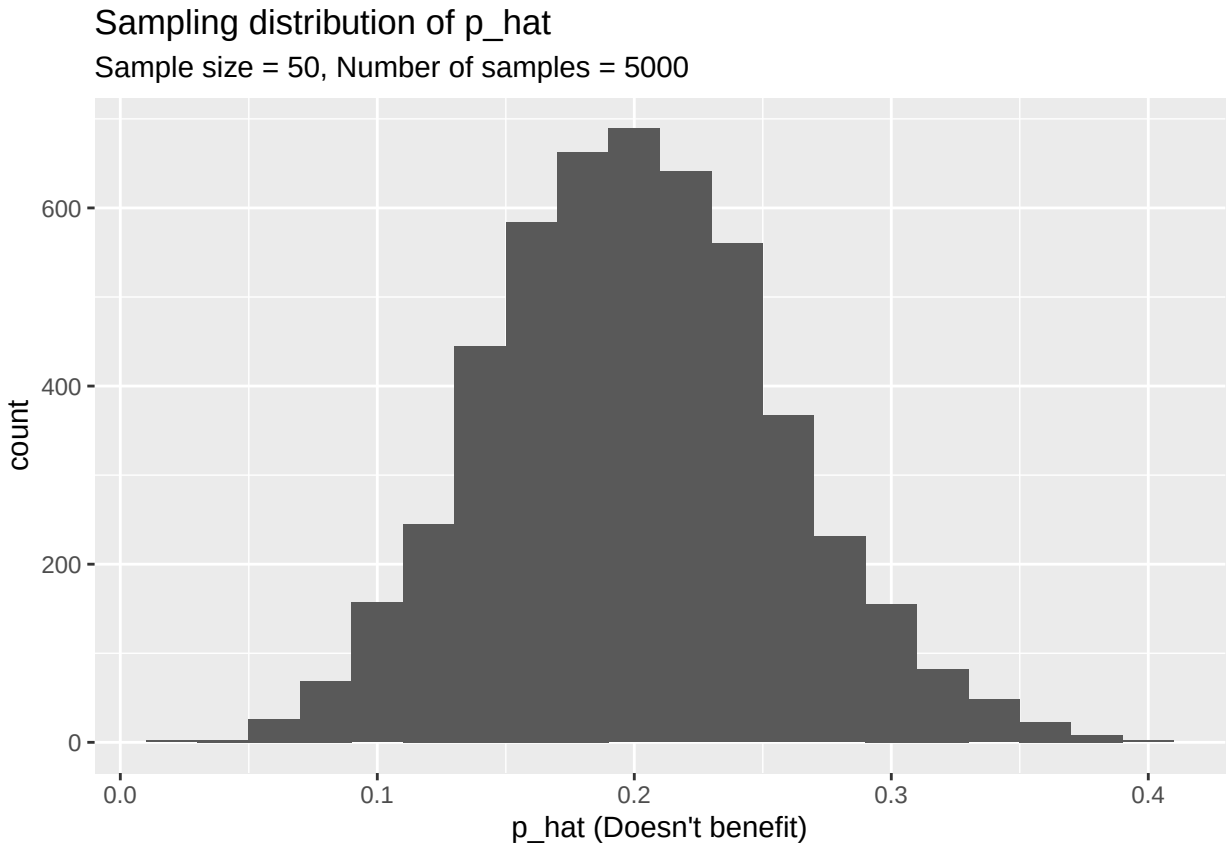
Each observation in the sampling distribution represent the estimating the true proportion of people who think that the work scientists do doesn't benefit them. As the sample size increases the mean of proportion gets closer to 0.2, When the sample size increases the se decreases because of less variability. The distribution looks normal when the sample size is larger because when the sample size is 10 the distribution is more of a right skew.




```r
sample_props50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 5000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
```

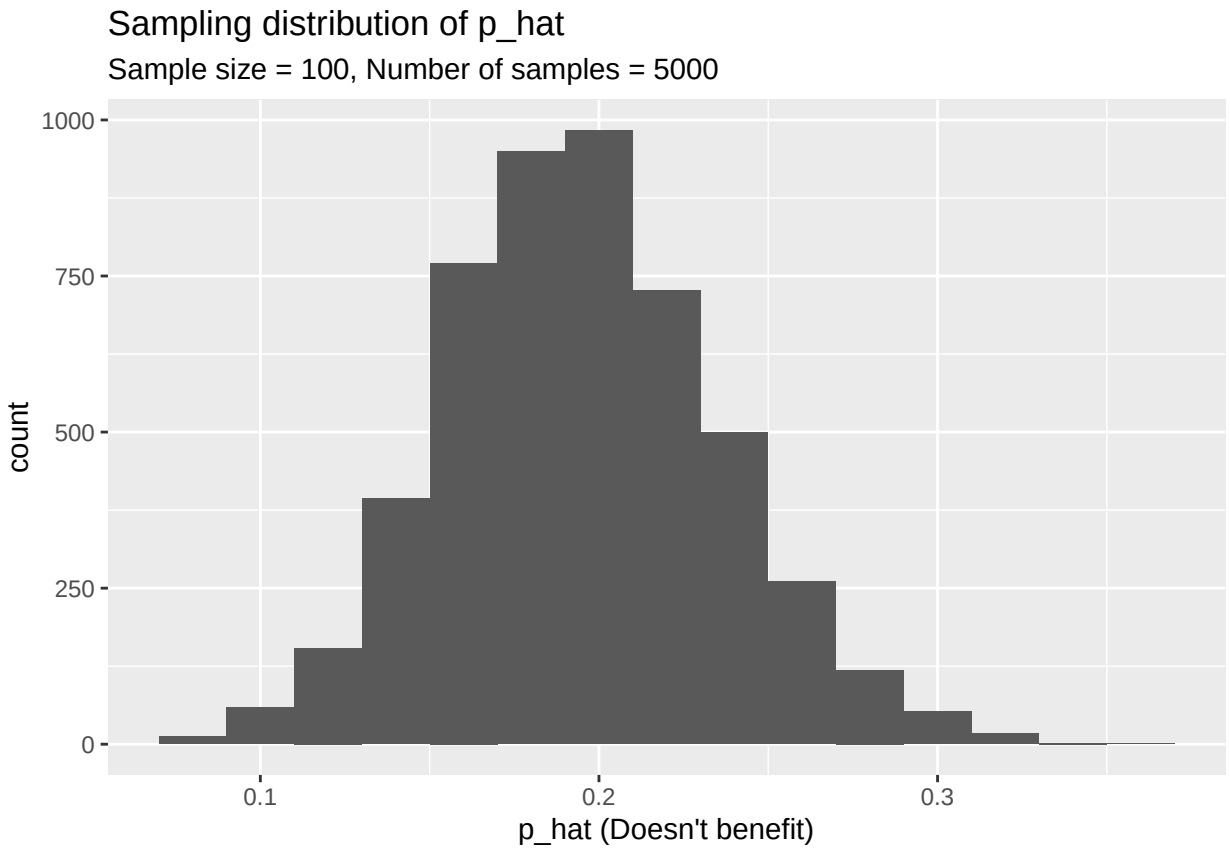
```
                    filter(scientist_work == "Doesn't benefit")

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 = 5000"
  )
```




```
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")

ggplot(data = sample_props100, 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 = 100, Number of samples = 5000"
  )
```



Exercise 7

```
set.seed(7)
global_monitor %>%
  sample_n(size = 15, replace = TRUE) %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Benefits")
```

The best point estimate of the population proportion of people who think the work scientists do enchances their lives is at 90%.

```
## # A tibble: 1 x 3
##   scientist_work     n p_hat
##   <chr>          <int> <dbl>
## 1 Benefits           9   0.6
```

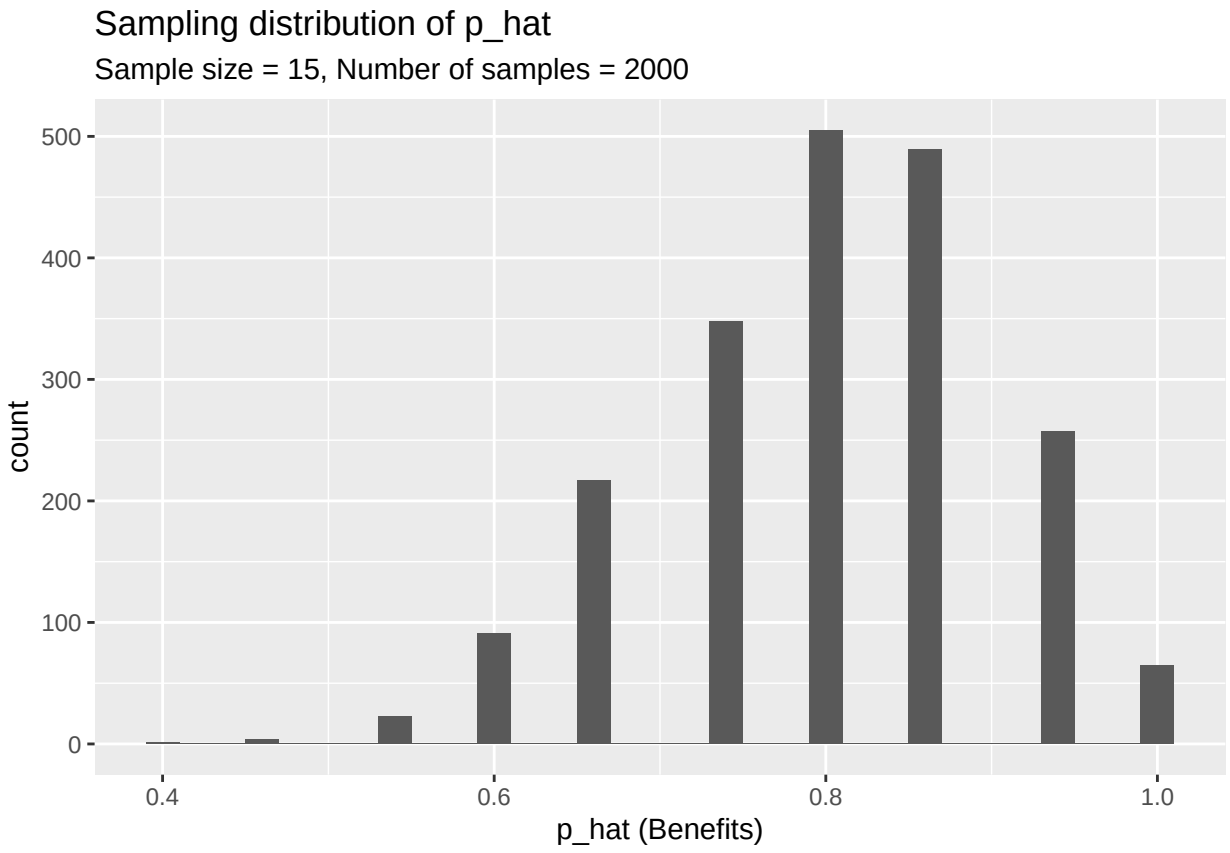
Exercise 8

```
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"
  )
```

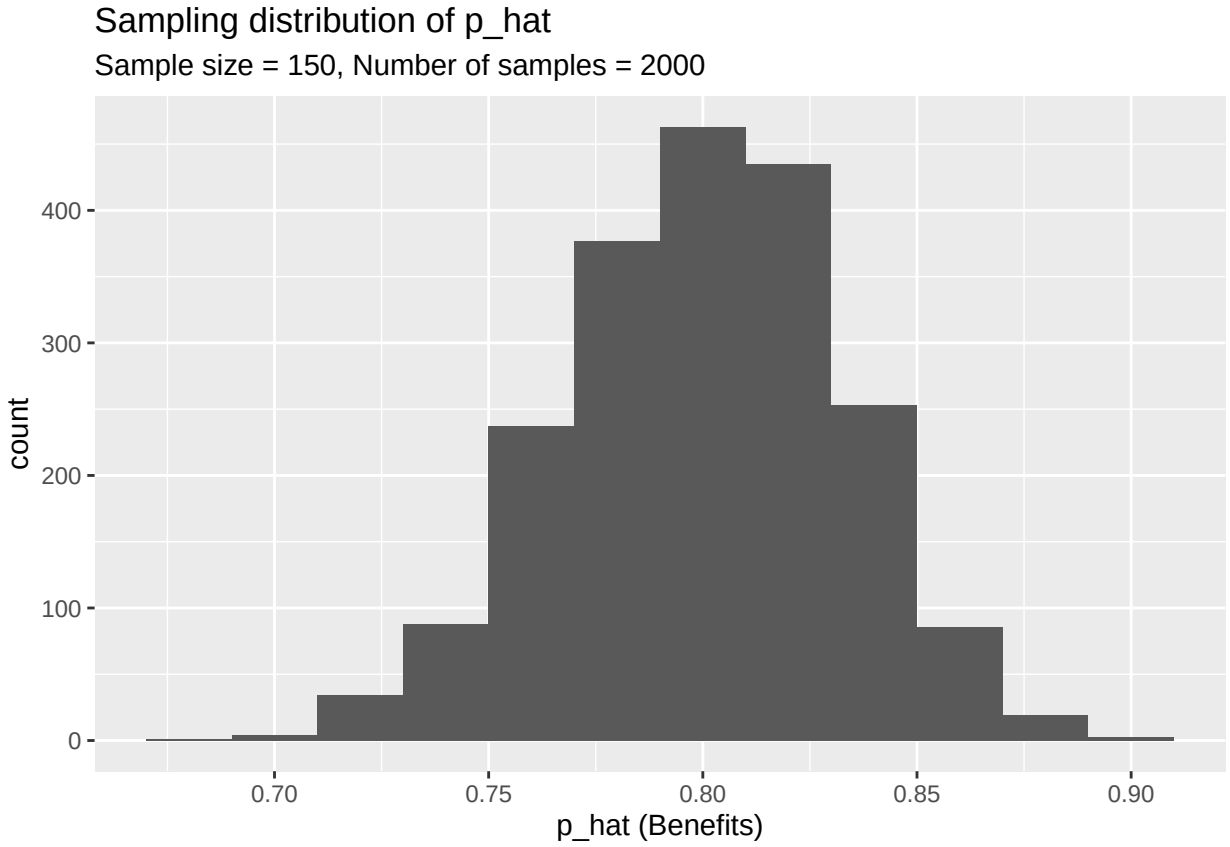
The shape of this sampling distribution is left skewed. The porportion seems far apart. Based on this sampling distribution, I would guess the true proportion of those who think the work scientists do enchances their lives to be is a approx. 250.




```
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"
  )
```



Exercise 9

The shape of this sampling distribution and compare to the sampling distribution of the sample size of 15 is kind of normal distribution but looks like a little left skewed. The proportion is gegting closer to the mean. Based on this sampling distribution, I would guess the true proportion of those who think the work scientists do enchances their lives to be is a littler over 750.

Exercise 10

Of the sampling distribution from 2 and 3 the more sample you add from 50 to 100, the more accurate is the measurement. Sample 3 has the smaller spread. The variance and the standard deviation are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value. In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. If I was concerned with making estimates that are more often close to the true value, I would prefer a sampling distribution with small spread becasue a small spread means that is more close to the true value.