DS606 - Lab5

Code

George Cruz

2020-10-04

Hide

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

global_monitor <- tibble(
  scientist_work = c(rep("Benefits", 80000), rep("Doesn't benefit", 20000))
)

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

Hide

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

```
## # A tibble: 2 x 3
##   scientist_work      n   pct
##   <chr>           <int> <dbl>
## 1 Benefits           35   0.7
## 2 Doesn't benefit    15   0.3
```

Based on the results, the distribution of responses in the sample is very similar to (in a case it was the same as) the distribution of responses in the population.

Exercise 2

I would expect the sample proportion to be very close to another student's sample proportion. There is a probability that they could match but I would expect the sample proportions to be distributed in a normal fashion.

Exercise 3

Hide

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

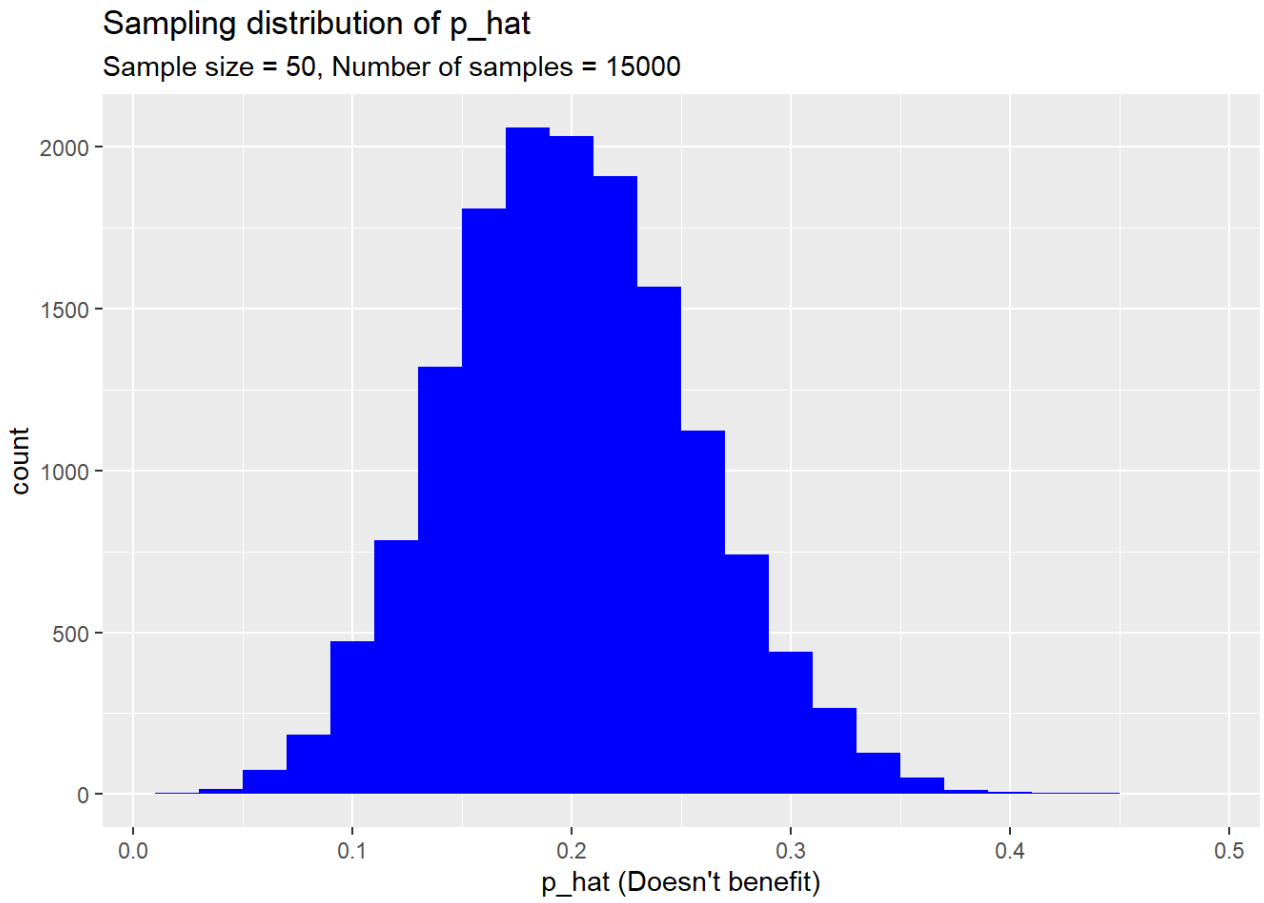
```
## # A tibble: 2 x 3
##   scientist_work      n   pct
##   <chr>           <int> <dbl>
## 1 Benefits           43  0.86
## 2 Doesn't benefit     7  0.14
```

Both proportions are close but not the same. If we took two other samples, one of size 100 and another of size 1000, I would expect the bigger sample proportions to be closer to the population's.

Hide

```
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, fill="blue") +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```



Exercise 4

There are 15000 samples in sampl50, the sample size for each is 50 and it shows a normal distribution, unimodal.

Exercise 5

Hide

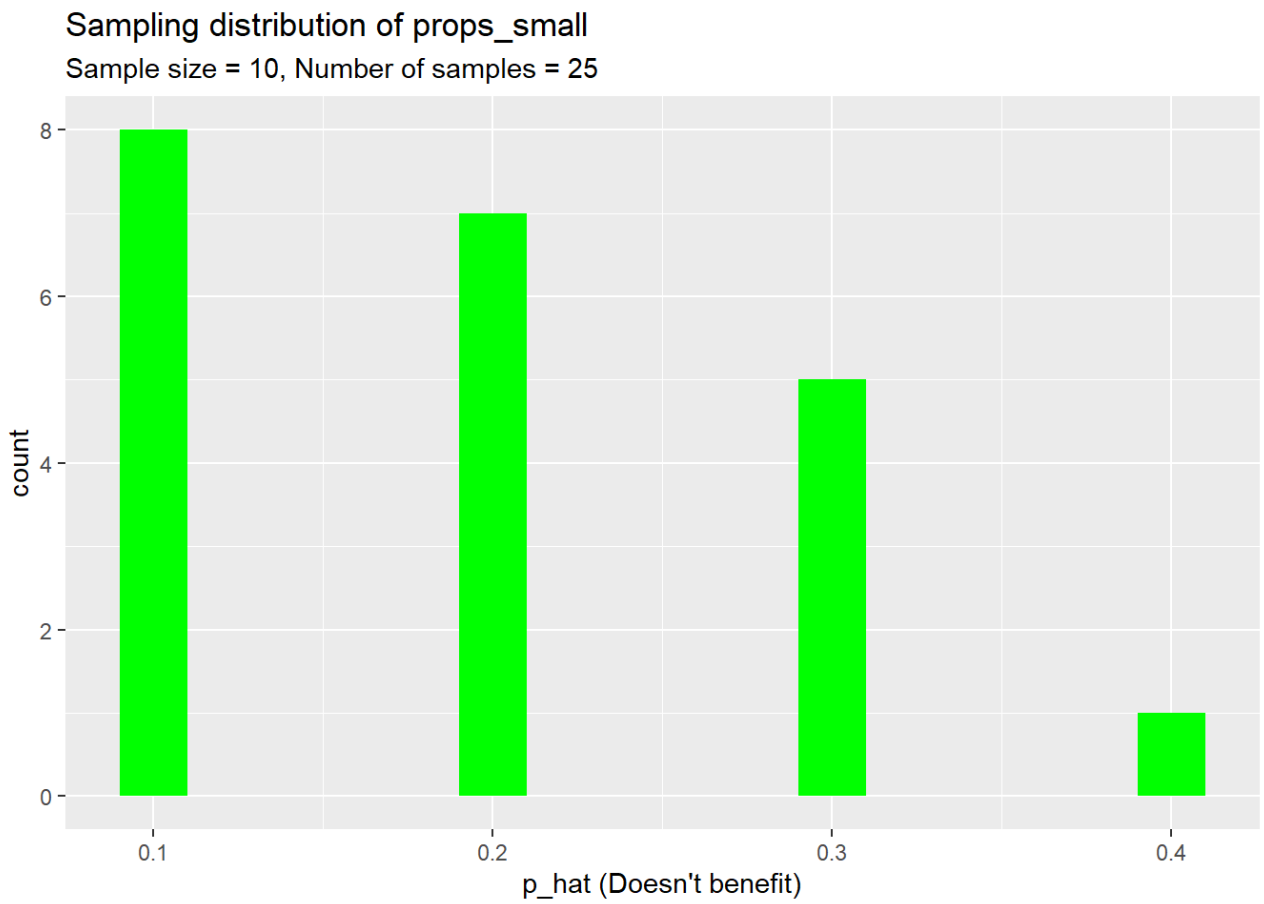
```
sample_props_small <- global_monitor %>%
                    rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

sample_props_small
```

```
## # A tibble: 21 x 4
## # Groups:   replicate [21]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     1   0.1
##  2         2 Doesn't benefit     4   0.4
##  3         3 Doesn't benefit     3   0.3
##  4         4 Doesn't benefit     1   0.1
##  5         5 Doesn't benefit     3   0.3
##  6         6 Doesn't benefit     3   0.3
##  7         8 Doesn't benefit     1   0.1
##  8         9 Doesn't benefit     2   0.2
##  9        11 Doesn't benefit     2   0.2
## 10        12 Doesn't benefit     2   0.2
## # ... with 11 more rows
```

Hide

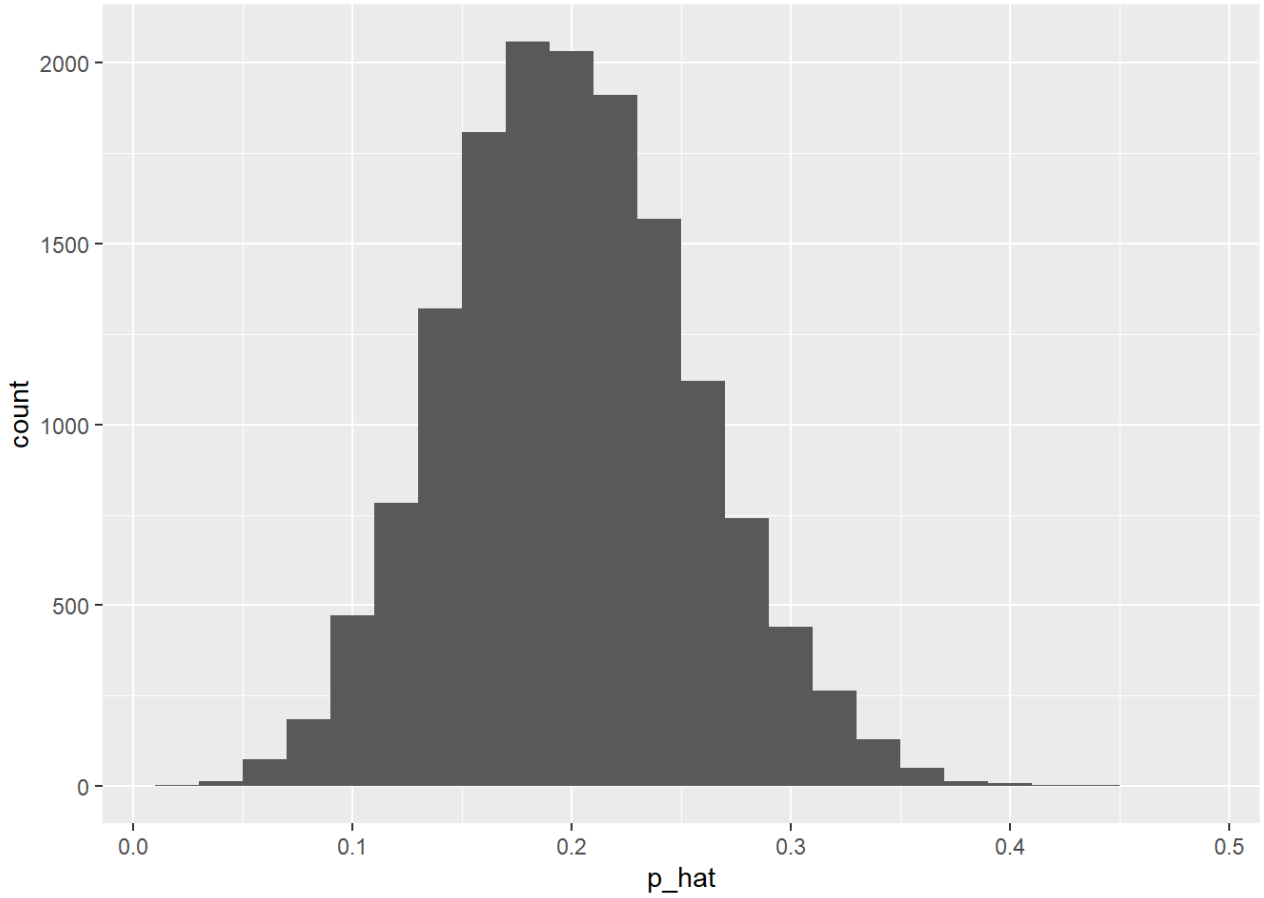
```
ggplot(data = sample_props_small, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02, fill="green") +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of props_small",
    subtitle = "Sample size = 10, Number of samples = 25"
  )
```



There are ~22 observations. Each observation represent a sample size of 10 values from the population.

Hide

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



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.)

Hide

```
library(plotrix)
sample_props <- 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")

values <- c( avg = mean(sample_props$p_hat), sd = sd(sample_props$p_hat, na.rm = TRUE
        ), se = std.error(sample_props$p_hat, na.rm = TRUE))
values
```

```
##         avg          sd          se
## 0.223213888 0.111582951 0.001664676
```

Hide

```
sample_props <- 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")

values <- c( avg = mean(sample_props$p_hat), sd = sd(sample_props$p_hat, na.rm = TRUE
         ), se = std.error(sample_props$p_hat, na.rm = TRUE))
values
```

```
##          avg           sd           se
## 0.1989040000 0.0559619972 0.0007914222
```

Hide

```
sample_props <- 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")

values <- c( avg = mean(sample_props$p_hat), sd = sd(sample_props$p_hat, na.rm = TRUE
         ), se = std.error(sample_props$p_hat, na.rm = TRUE))
values
```

```
##          avg           sd           se
## 0.1999720000 0.0400674468 0.0005666393
```

As we increase the sample size, the mean gets closer to 0.20 and the standard error gets smaller, approaching 0.

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 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?

Hide

```
set.seed(50)
sampl_15 <- global_monitor %>%
  sample_n(15) %>%
  count(scientist_work) %>%
  mutate(pct = n /sum(n))

sampl_15
```

```
## # A tibble: 2 x 3
##   scientist_work      n   pct
##   <chr>           <int> <dbl>
## 1 Benefits           11 0.733
## 2 Doesn't benefit     4 0.267
```

Using these values, I would estimate that 73% of the population think the work scientists do enhances their lives.

Exercise 8

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 sample_props15.

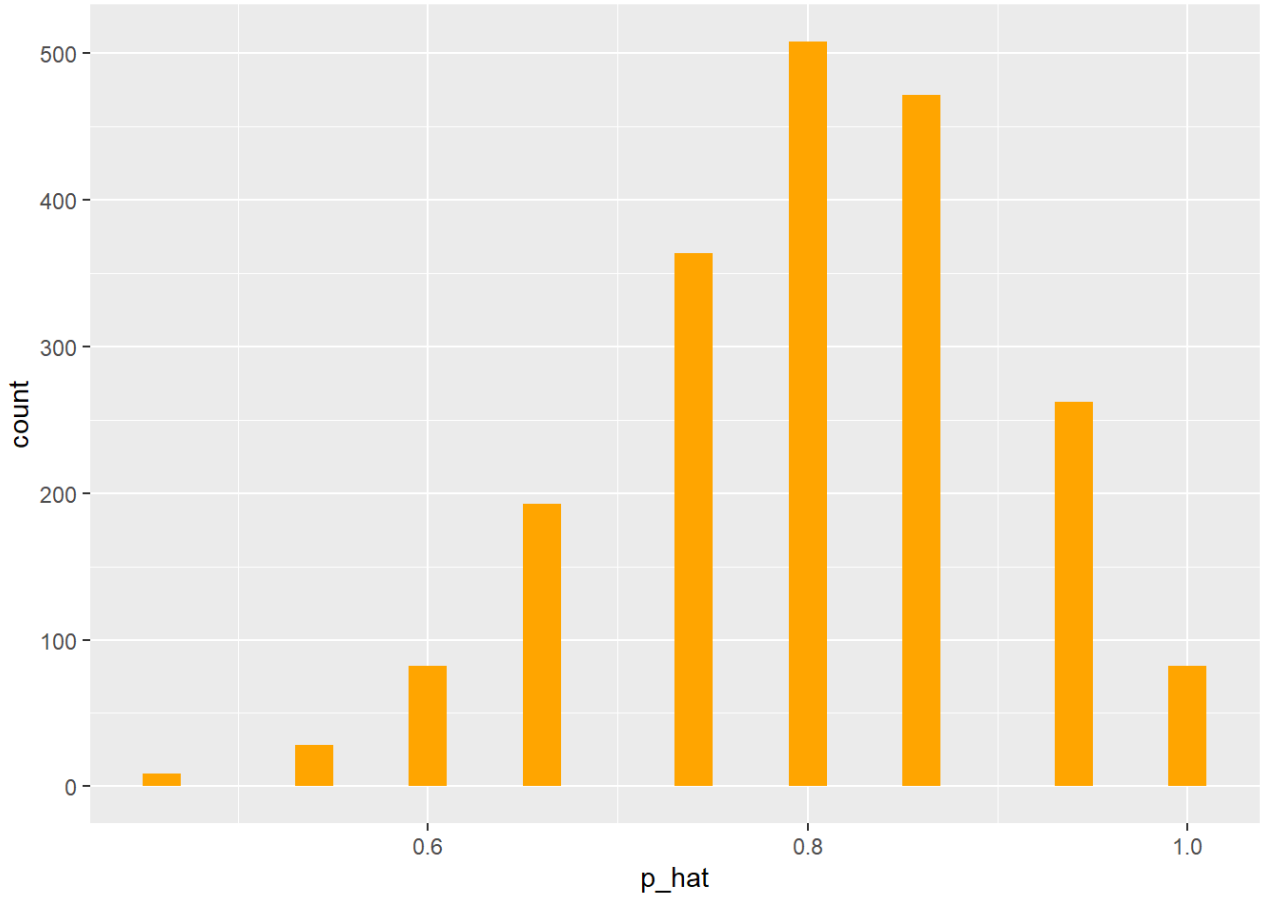
Hide

```
set.seed(Sys.time())  #remove previous seed
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")
```

Plot the data, then describe the shape of this sampling distribution.

Hide

```
ggplot(data = sample_props15, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02, fill="orange")
```



The resulting histogram is left skewed, unimodal and approximates a normal 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.

By looking at this sampling distribution, I would guess 80% is the true proportion.

Hide

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

```
## # A tibble: 2 x 3
##   scientist_work      n   pct
##   <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.

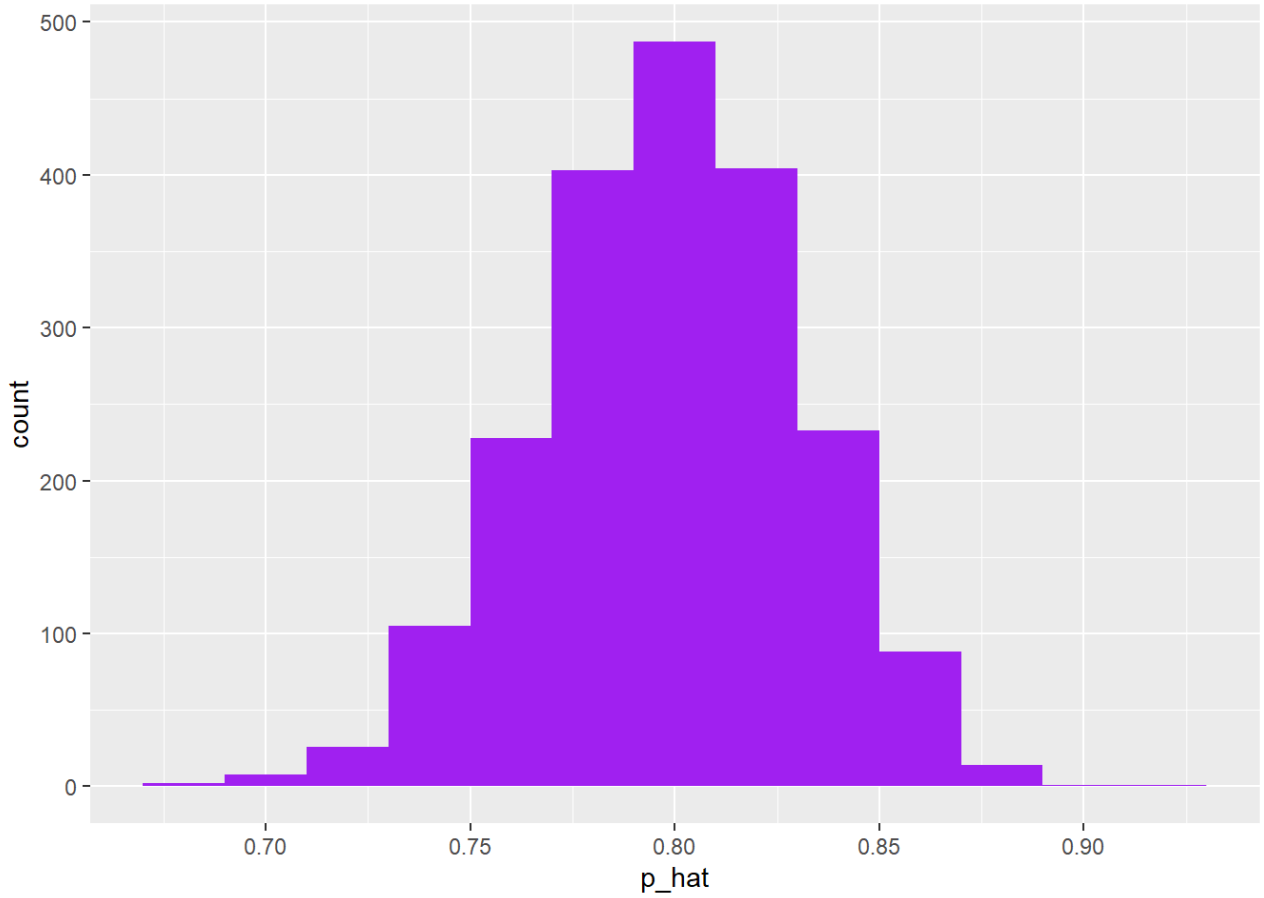
Hide

```
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")
```

Describe the shape of this sampling distribution and compare it to the sampling distribution for a sample size of 15.

Hide

```
ggplot(data = sample_props150, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02, fill="purple")
```



This distribution is normal, unimodal.

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?

Around 80%

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 bigger sample size has a smaller spread. I would prefer a sampling distribution with a small spread.

…