Sampling Distributions

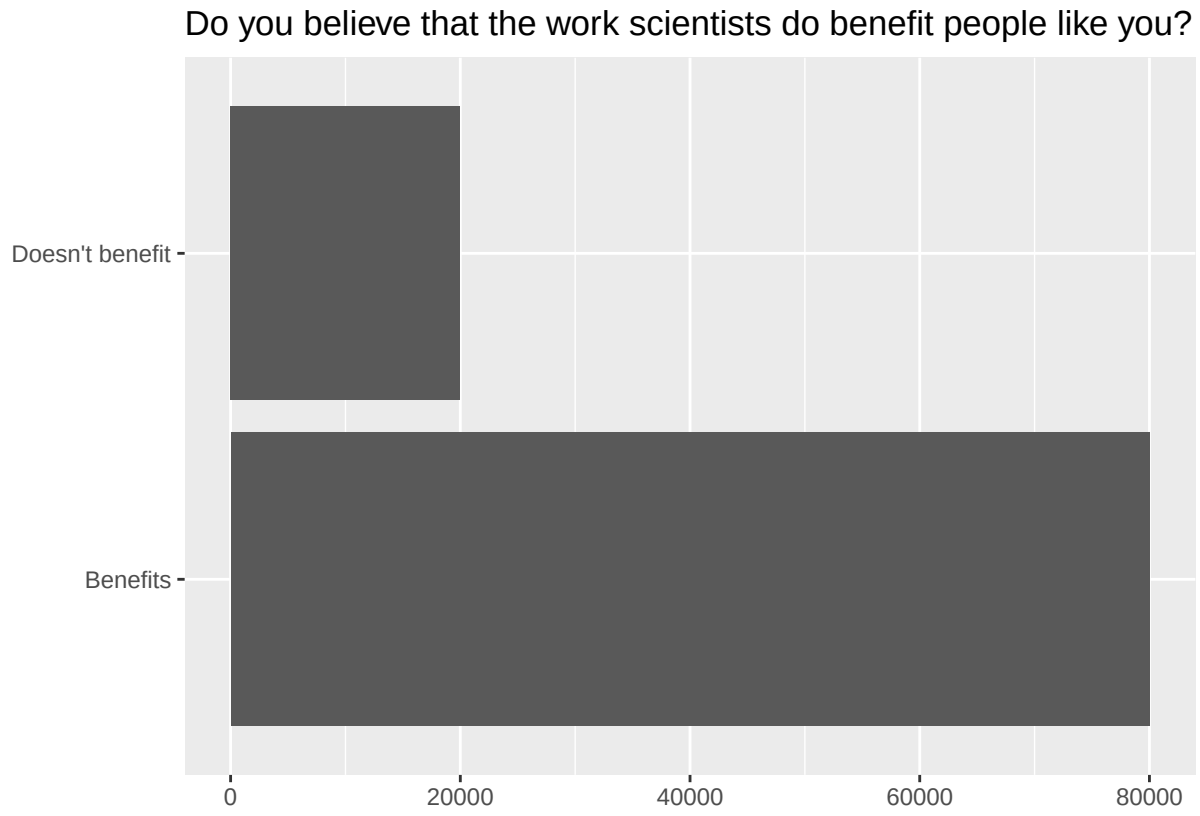
Zachary Palmore

2020-10-04

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

Pre-exercise

```
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()
```



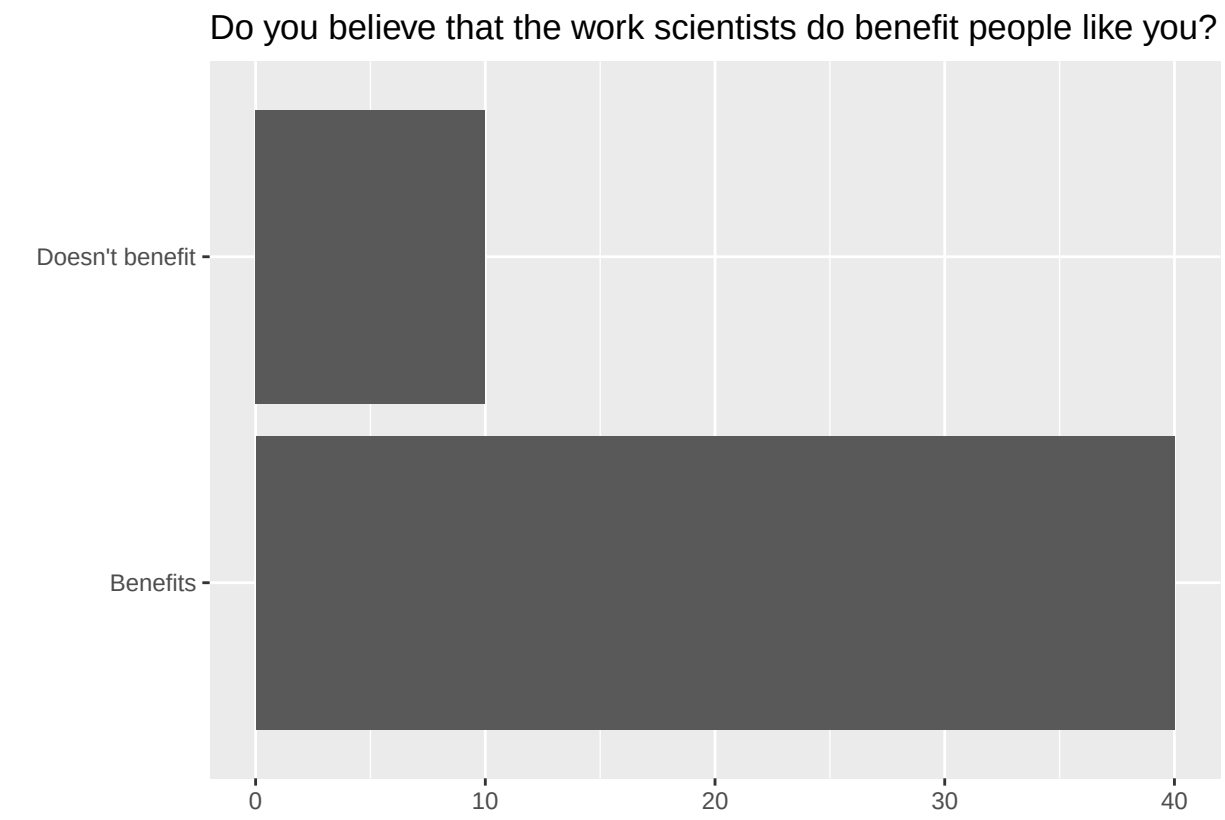

```
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

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.

Reusing the code to generate a sample, observing its distribution and proportions.

```
set.seed(09292020)
samp1 <- global_monitor %>%
  sample_n(50)
View(samp1)
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```




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

```
## # A tibble: 2 x 3
##   scientist_work      n sample_proportion_phat
##   <chr>           <int>                  <dbl>
## 1 Benefits           40                    0.8
## 2 Doesn't benefit    10                    0.2
```

In this case, the distribution of sample proportions is the same as the global monitor's proportions. About 80% (or 0.8) of respondents in both believe that the work scientists do is helpful to them, while 20% percent (or 0.2) do not.

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.

The sample proportion of another student's sample can be generated by repeating the process. For good measure, I will run it with 5 "other students."

```r
# Student 1
set.seed(19312020)
student1_samp1 <- global_monitor %>%
  sample_n(50)
student1_samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # 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
```

In this first run, the proportions were close to matching but not identical to that of my own analysis. Importantly, the values were selected at random for the new sample.

```r
# Student 2
set.seed(29312020)
student2_samp1 <- global_monitor %>%
  sample_n(50)
student2_samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           40   0.8
## 2 Doesn't benefit    10   0.2
```

In this second run, the proportions were identical to my original run. This is again using new, randomly selected values from the global_monitor data to create the sample.

```r
# Student 3
set.seed(39312020)
student3_samp1 <- global_monitor %>%
  sample_n(50)
student3_samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # 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
```

Third, we get a match to the first student's run but not identical to my original.

```
# Student 4
set.seed(49312020)
student4_samp1 <- global_monitor %>%
  sample_n(50)
student4_samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           40   0.8
## 2 Doesn't benefit    10   0.2
```

Fourth is another identical match to my samp1 or the original population proportions.

```
# Student 5
set.seed(59312020)
student5_samp1 <- global_monitor %>%
  sample_n(50)
student5_samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           40   0.8
## 2 Doesn't benefit    10   0.2
```

Lastly, in this fifth run we have another exact match to the population proportion and samp1.

Although the population proportion and samp1 proportion generated gave the proportions 0.8 and 0.2, not all other students (of the 5 other random student#_samp1) had the same results. Therefore, in some cases the proportions will match, while in other cases the sample will be slightly different but never too far from the population when sampling from the same data.

In short, we should expect similar results but not identical ones. This is because we are selecting the values from the data at random, which will produce variations in the proportions. However, the proportions should be similar because it is selected from the same data source (or population).

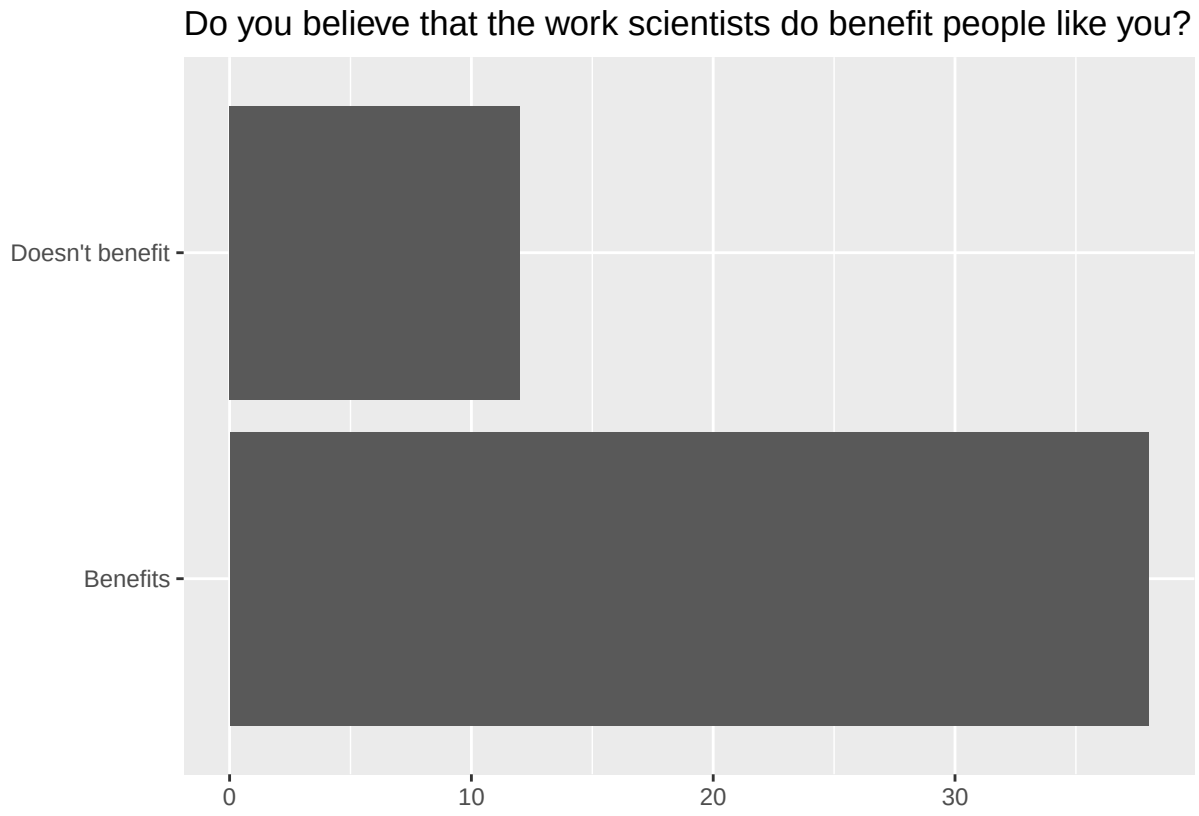
Exercise 3

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?

Running sample 2 chunk to collect a sample from the global monitor data and calculate its proportions.

```
set.seed(09302020)
samp2 <- global_monitor %>%
  sample_n(50)
```

```
View(samp2)
ggplot(samp2, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```




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

```
## # A tibble: 2 x 3
##   scientist_work      n sample_proportion
##   <chr>           <int>             <dbl>
## 1 Benefits           38              0.76
## 2 Doesn't benefit    12              0.24
```

Results show that the sample proportions of samp2 at 0.76 and 0.24 are very similar to that of samp1 at 0.80 and 0.20. While samp2 is slightly lower than samp1, two of the 'other five students' had proportions that were slightly higher than samp1. If we ran this repeatedly, the results should remain similar.

Now, lets take two more samples and find their proportions. One will be of the sample size 100 and the other 1000. All other parameters will remain the same.

```r
# Sample size of 100
set.seed(09322020)
samp_100 <- global_monitor %>%
  sample_n(100)
View(samp_100)
samp_100 %>%
  count(scientist_work) %>%
  mutate(sample_proportion = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n sample_proportion
##   <chr>           <int>             <dbl>
## 1 Benefits           85              0.85
## 2 Doesn't benefit    15              0.15
```

The proportions of sample size 100 are 0.85 that believe what scientists do benefits them to 0.15 that do not believe what scientist do benefits them. These proportions are very similar to the original population or samp1. They are also close to the results of the other simulations.

```r
# Sample size of 1000
set.seed(09332020)
samp_1000 <- global_monitor %>%
  sample_n(1000)
View(samp_1000)
samp_1000 %>%
  count(scientist_work) %>%
  mutate(sample_proportion = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n sample_proportion
##   <chr>           <int>             <dbl>
## 1 Benefits          797             0.797
## 2 Doesn't benefit   203             0.203
```

With the sample size at 1000 the proportions were even closer to the that of the original population. To be sure that this trend continues, I will run 5 more samples at this size.

```r
# Other sample #1 size of 1000
set.seed(09342020)
samp1_1000 <- global_monitor %>%
  sample_n(1000)
samp1_1000 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          805 0.805
## 2 Doesn't benefit   195 0.195
```

```r
# Other sample #2 size of 1000
set.seed(09352020)
samp2_1000 <- global_monitor %>%
  sample_n(1000)
samp2_1000 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          815 0.815
## 2 Doesn't benefit   185 0.185
```

```r
# Other sample #3 size of 1000
set.seed(09362020)
samp3_1000 <- global_monitor %>%
  sample_n(1000)
samp3_1000 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          801 0.801
## 2 Doesn't benefit   199 0.199
```

```r
# Other sample #4 size of 1000
set.seed(09372020)
samp4_1000 <- global_monitor %>%
  sample_n(1000)
samp4_1000 %>%
  count(scientist_work) %>%
  mutate(p_hat= n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          798 0.798
## 2 Doesn't benefit   202 0.202
```

```r
# Other sample #5 size of 1000
set.seed(09382020)
samp5_1000 <- global_monitor %>%
  sample_n(1000)
samp5_1000 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
```

```
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          810  0.81
## 2 Doesn't benefit   190  0.19
```

Based on these six sample proportions (which includes the first sampling set at seed 09332020), it appears having a larger sample size gives a slightly closer estimate to the proportions of the entire population from the global monitor data. Therefore, having a larger sample size (in this case from 100 to 1000) seems to increase the accuracy in estimating the population proportions by reducing the sampling error.

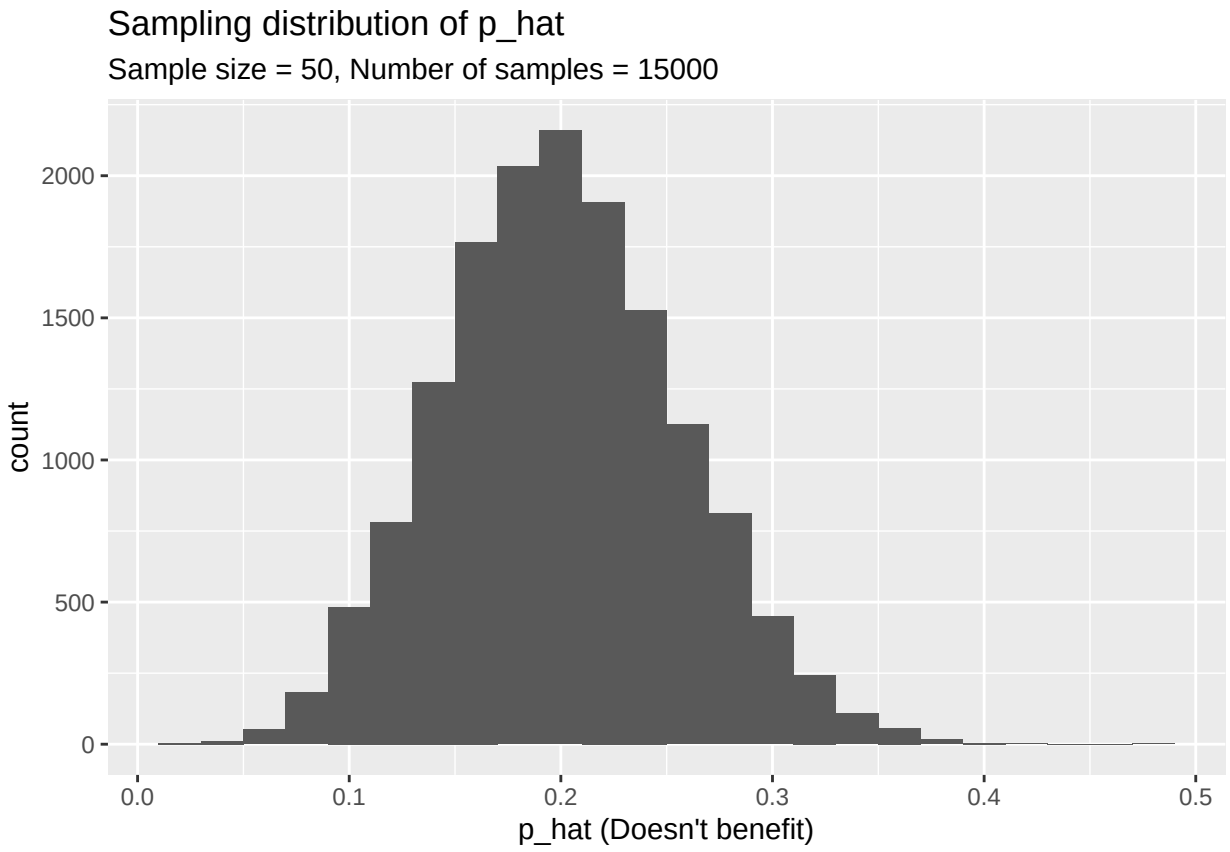
However, simply having a larger sample size may not always result in closer estimates. For example, the samp1 and several 'other students' samples gave better (exact) estimations of the entire population. Yet their sample size was smaller.

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.

Creating the sample_props50 using the given lab code and visualizing.

```
set.seed(09272020)
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"
  )
```



There are 15000 elements of 4 variables in sample_props50. Its is centered at 0.20 with a symmetric shape and relatively even distribution of data spreading out from its center. It only has the one peak at 0.20, making it unimodal, and the data thins out before reaching 0.0 (at the lower end) and 0.45 (at the higher end). The center is located at the population proportion.

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?

Each observation represents the proportion of "Doesn't benefit" responses of 10 randomly selected individuals from the total population in the global monitor data. Using the code from lab we can try to alter the chunk to create this;

```
# Where sample size = 10 and  25 samples (reps)
set.seed(09612020)
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")
glimpse(sample_props_small)
```

```
## Rows: 24
```

```
## Columns: 4
## Groups: replicate [24]
## $ replicate      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ scientist_work <chr> "Doesn't benefit", "Doesn't benefit", "Doesn't benef...
## $ n              <int> 3, 2, 4, 3, 2, 4, 2, 3, 3, 2, 1, 1, 2, 3, 2, 3, 1, 3...
## $ p_hat          <dbl> 0.3, 0.2, 0.4, 0.3, 0.2, 0.4, 0.2, 0.3, 0.3, 0.2, 0....
```

However, this results in a different number of reps than specified. In this case, there are 24 observations of 4 variables. We can see that it is missing the 19th sample. Why are there not exactly 25 as specified? First, let's run it again to see if this continues.

```
# Where sample size = 10 and for 25 random samples
set.seed(09622020)
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")
glimpse(sample_props_small)
```

```
## Rows: 20
## Columns: 4
## Groups: replicate [20]
## $ replicate      <int> 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ scientist_work <chr> "Doesn't benefit", "Doesn't benefit", "Doesn't benef...
## $ n              <int> 2, 3, 2, 2, 2, 1, 2, 2, 2, 3, 1, 2, 2, 2, 3, 2, 1, 4...
## $ p_hat          <dbl> 0.2, 0.3, 0.2, 0.2, 0.2, 0.1, 0.2, 0.2, 0.2, 0.3, 0....
```

This time there are only 20 observations of 4 variables. How could this be?

Consider the success-failure condition of the Central Limit Theorem (CLT). For the theorem to hold, and the sample produce a normal distribution, observations must be independent and the sample size must be large enough to do so. The rule is that the product of the sample size and population proportion must be greater than or equal to 10 and the the product of the sample size and one minus the population proportion must also be greater than or equal to 10. We can test to see if these values hold up the theorem.

```
# Where sample size = 10 and the population proportion is 0.20
n <- 10
p <- 0.20
n*p
```

```
## [1] 2
```

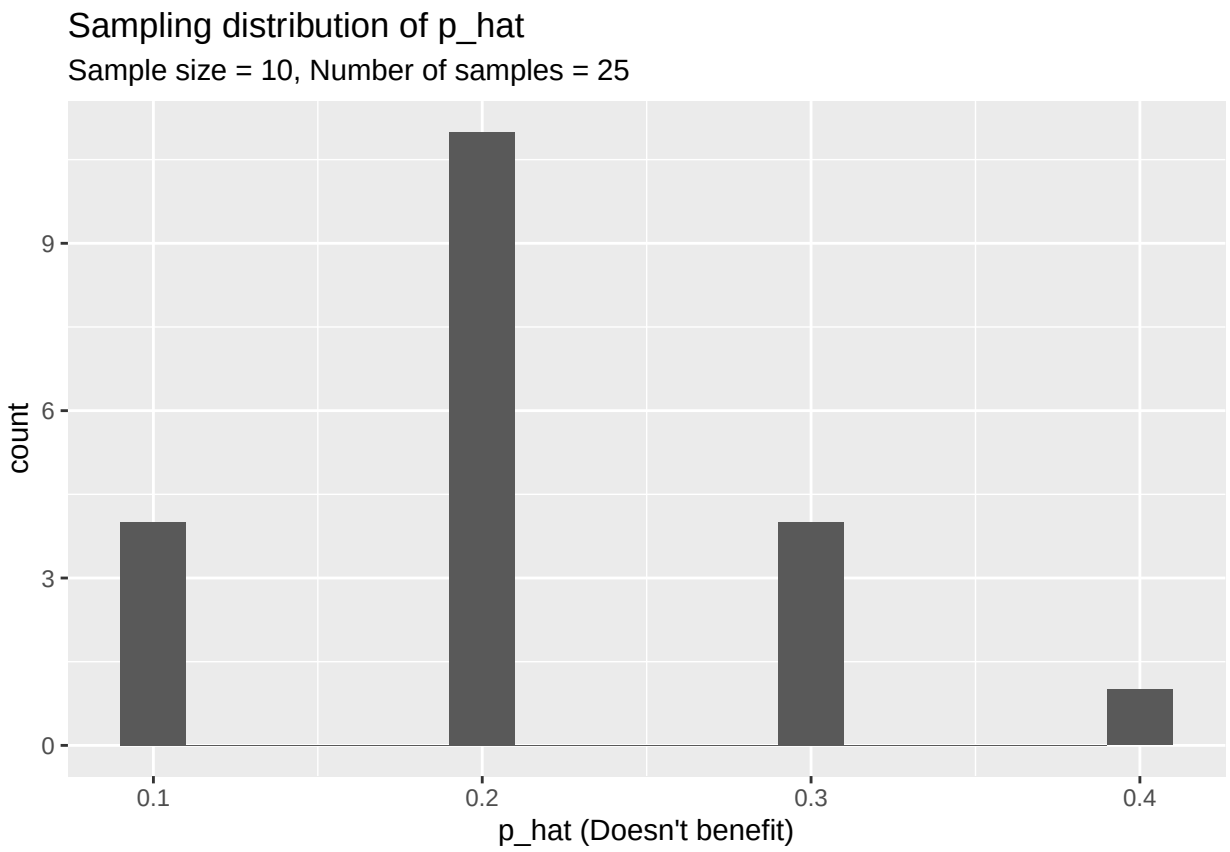
```
n*(1-p)
```

```
## [1] 8
```

We get the results of $np = 2$ and $n(1-p) = 8$. They do not support the Central Limit Theorem because they are not greater than or equal to 10. Since we know the observations are independent, this indicates that at this sample size, a normal distribution might not be appropriate. The sample size would need to be increased to generate a more normal distribution.

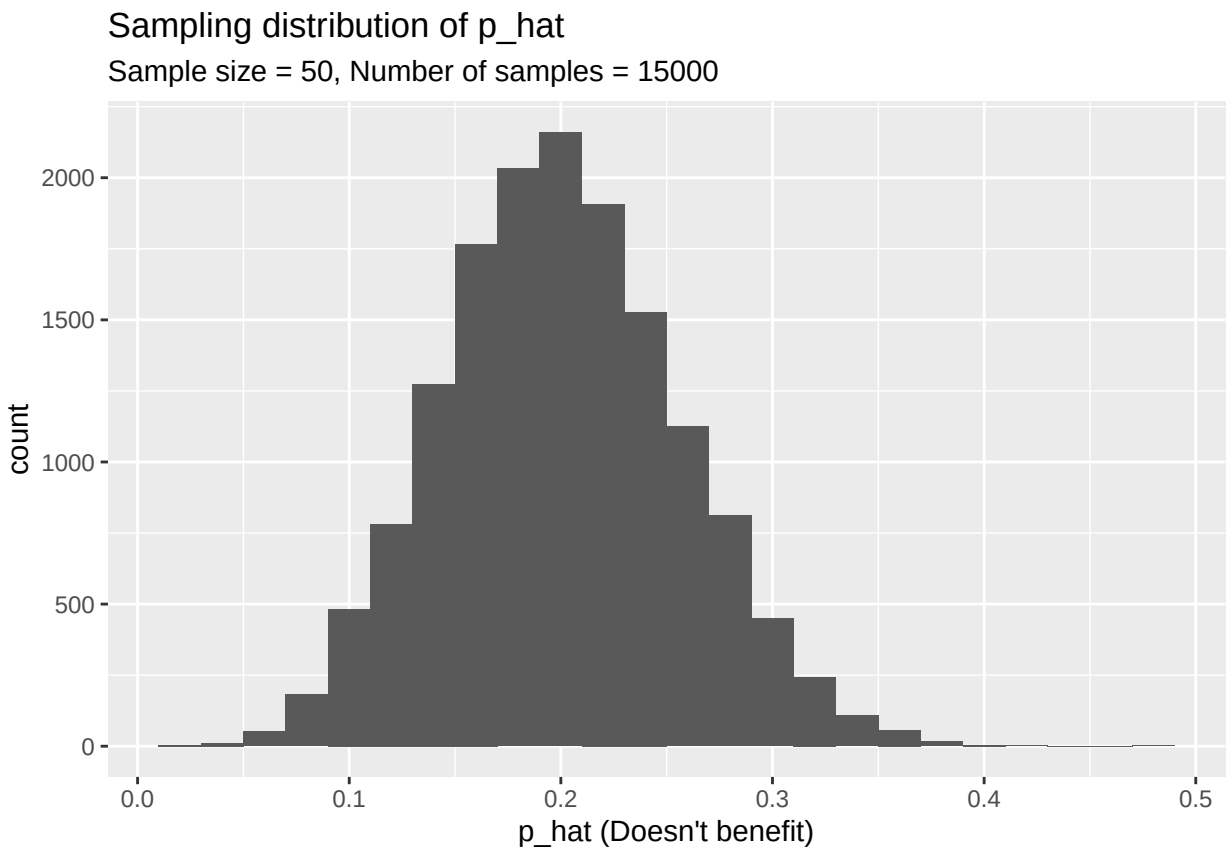
For visualizing what the most recent small sample size looks like we have this:

```r
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" )
```



It is easy to point out a few distinct features of this distribution. First, the data is more discrete than a larger sample. For example in a sample size of 50 with 15000 samples (plotted below), we can see a more continuous distribution across the x-axis.

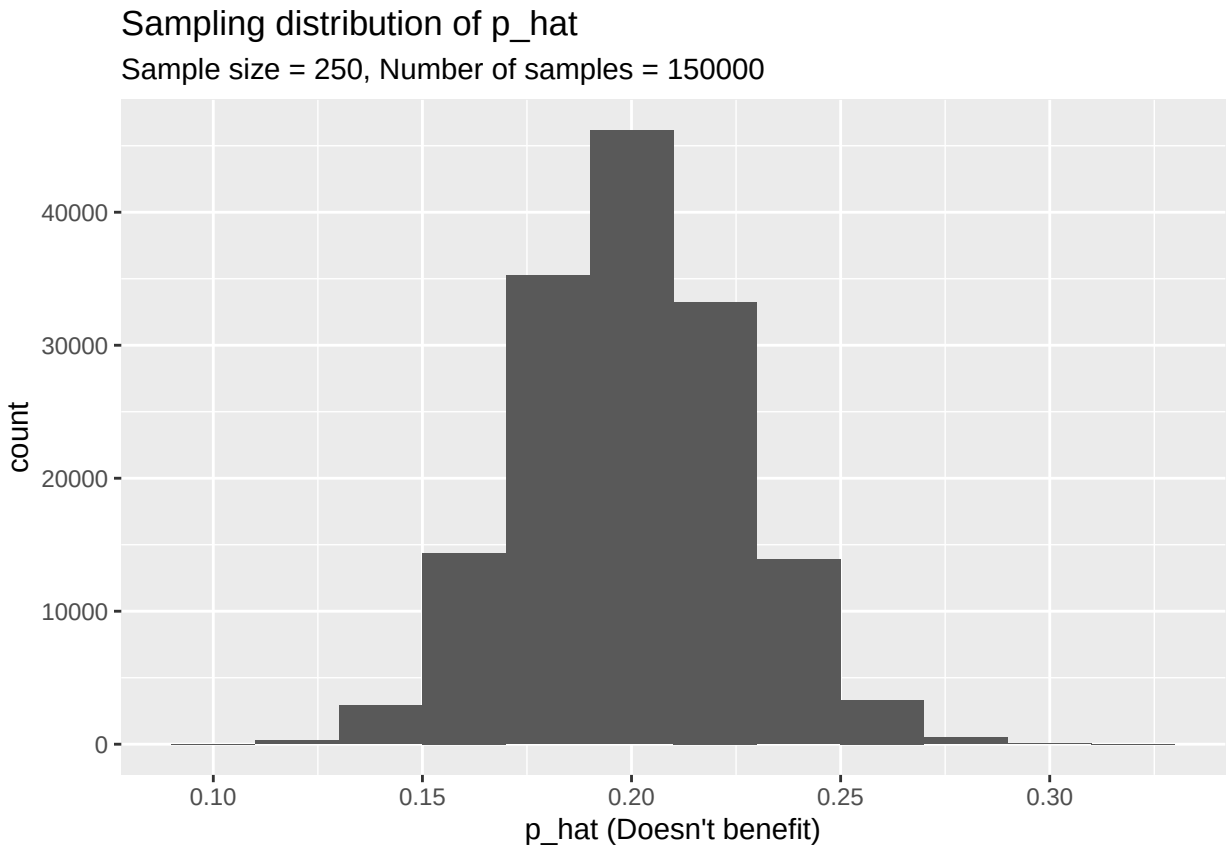
```r
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```



Data fills the space between the values at this larger sample size.

Another feature is the skew of a smaller sample size. In the first sample size 10, the points at 0.40 cause the distribution to appear left of center. With larger sample sizes, this is not as noteworthy. For example, we can make the sample size 500 at 15000 samples.

```
set.seed(09012020)
sample_example_large <- global_monitor %>%
                    rep_sample_n(size = 250, reps = 150000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
ggplot(data = sample_example_large, 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 = 250, Number of samples = 150000"
  )
```



In this distribution data appears more symmetric, it is still centered at the population parameter but the x-axis appears to have shrunk. This follows the CLT that the larger the sample size, the more normal the distribution will appear to be. Having a larger sample size will also make the distribution appear more continuous.

Exercise 6

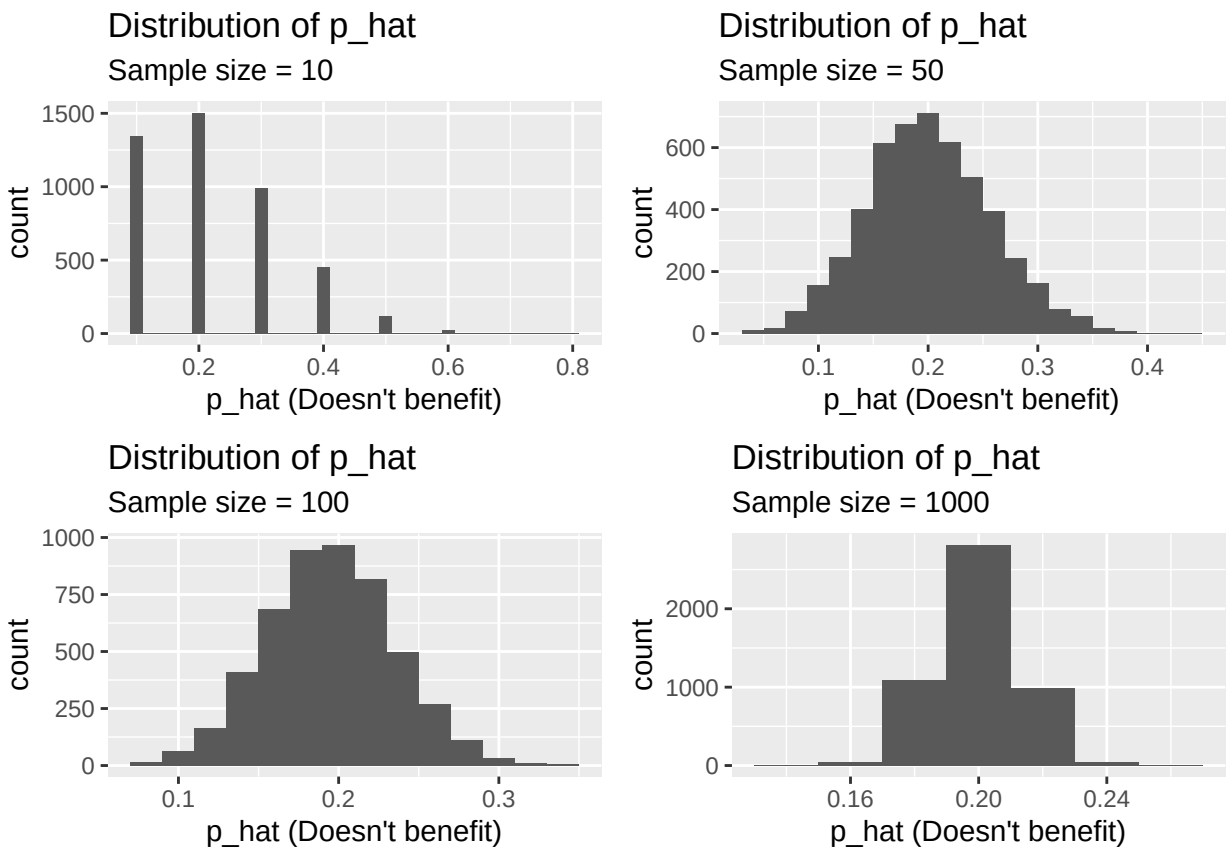
Use the app below to create sampling distributions of proportions of Doesn't benefit from samples of size 10, 50, and 100. Use 5,000 simulations. What does each observation in the sampling distribution represent? How does the mean, standar error, and shape of the sampling distribution change as the sample 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.)

Creating distributions of proportions of "Doesn't benefit" from samples of size 10, 50, and 100 using 5,000 simulations each.

```
# At sample size 10
set.seed(09022020)
sample_size10 <- 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")
# At sample size 50
set.seed(09032020)
sample_size50 <- 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")
# At sample size 100
set.seed(09082020)
sample_size100 <- 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")
# At sample size 1000
set.seed(09052020)
sample_size1000 <- global_monitor %>%
                    rep_sample_n(size = 1000, reps = 5000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
ggsize10 <- ggplot(data = sample_size10, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 10"
  )
ggsize50 <- ggplot(data = sample_size50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 50"
  )
ggsize100 <- ggplot(data = sample_size100, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 100"
  )
ggsize100 <- ggplot(data = sample_size100, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 100"
  )
ggsize1000 <- ggplot(data = sample_size1000, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 1000"
  )
```

```
ggarrange(ggsize10, ggsize50, ggsize100, ggsize1000)
```



Each observation in the sampling distribution represents a randomly selected set of responses from the global monitor data that were of the belief that science doesn't benefit them individually. In each distribution, the mean is the same at 0.20. This can be validated with the mean and standard error (SE) calculations of each sample size in the data frame.

```
SampleStats <- as.data.frame(c(10, 50, 100, 1000))
SampleStats$Size <- SampleStats$`c(10, 50, 100, 1000)`
SampleStats$Reps <- c(5000)
SampleStats$Mean <- c((mean(sample_size10$p_hat)),(mean(sample_size50$p_hat)),(mean(sample_size100$p_ha
# Function for standard error
st_error <- function(n, p) {
  mth <- p*(1-p)
  SE <- sqrt(mth/n)
  return(SE)
}
SampleStats$p <- c(0.20)
SampleStats$SE <- st_error(SampleStats$Size, SampleStats$p)
SampleStats <- SampleStats[,2:6]
SampleStats
```

```
##   Size Reps      Mean   p         SE
## 1   10 5000 0.2230527 0.2 0.12649111
## 2   50 5000 0.2007640 0.2 0.05656854
```

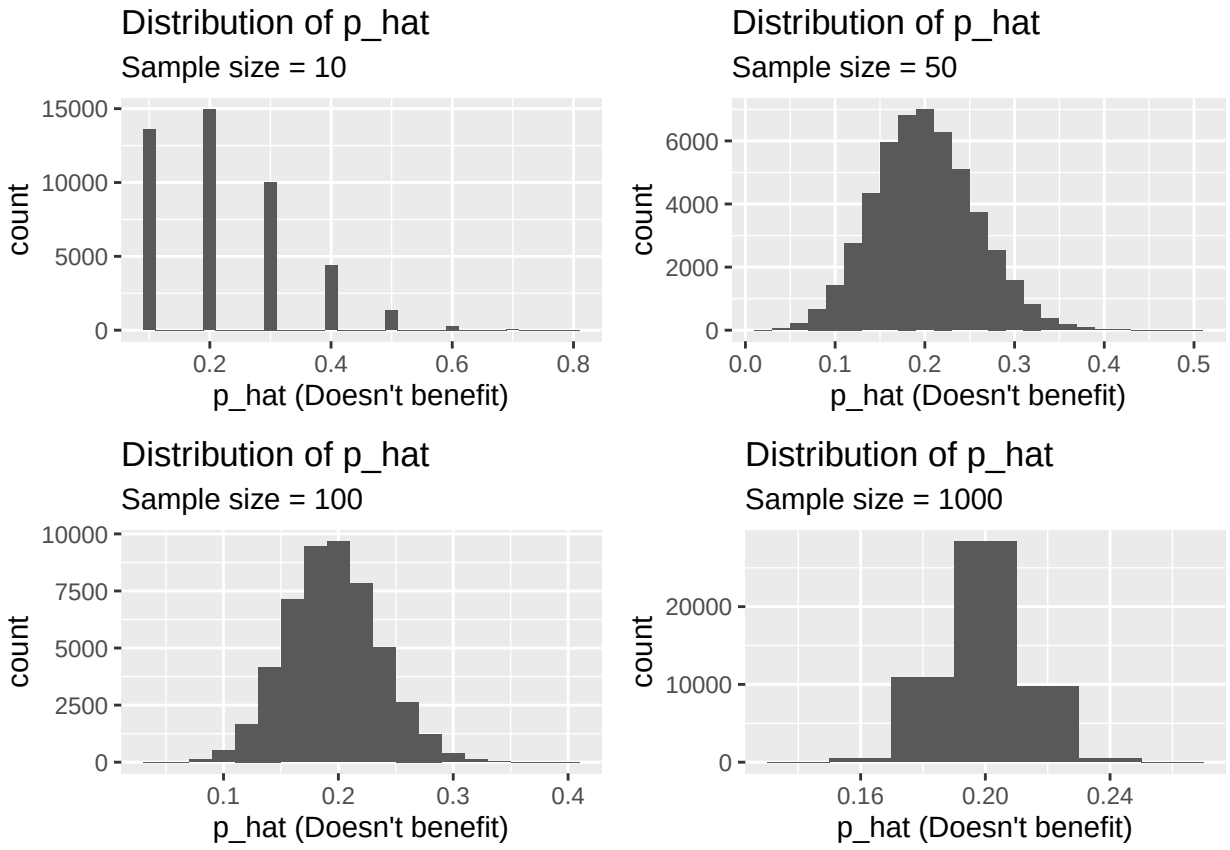
```
## 3  100 5000 0.1999580 0.2 0.04000000
## 4 1000 5000 0.1999350 0.2 0.01264911
```

The standard error decreases as the sample size increases and the shape of the sampling distribution also narrows with an increase in sample size. The mean also gets closer to the population proportion (p).

To understand how the number of simulations effects the mean and standard error we can add new samples with more simulations to the data frame. If there are any larger changes in the numbers, then we should see. While doing so, we can also visualize the results.

```
# At sample size 10
set.seed(09102020)
sample2_size10 <- global_monitor %>%
                    rep_sample_n(size = 10, reps = 50000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
# At sample size 50
set.seed(09112020)
sample2_size50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 50000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
# At sample size 100
set.seed(09122020)
sample2_size100 <- global_monitor %>%
                    rep_sample_n(size = 100, reps = 50000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
# At sample size 1000
set.seed(09132020)
sample2_size1000 <- global_monitor %>%
                    rep_sample_n(size = 1000, reps = 50000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
gg2size10 <- ggplot(data = sample2_size10, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 10"
  )
gg2size50 <- ggplot(data = sample2_size50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 50"
  )
gg2size100 <- ggplot(data = sample2_size100, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
```

```
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 100"
  )
gg2size1000 <- ggplot(data = sample2_size1000, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Distribution of p_hat",
    subtitle = "Sample size = 1000"
  )
ggarrange(gg2size10, gg2size50, gg2size100, gg2size1000)
```



At first glance we can see that at higher numbers of simulations, there is an increase of counts on the y-axis. This is good, because we added samples by changing its quantity from 5,000 to 50,000. The variations that occur with smaller sample sizes are still apparent. For example, at a sample size of 10, we can see the tail of this data with higher reps is quite the same as the first distribution with fewer reps.

```
SampleStats2 <- as.data.frame(c(10, 50, 100, 1000))
SampleStats2$Size <- SampleStats2$`c(10, 50, 100, 1000)`
SampleStats2$Reps <- c(50000)
SampleStats2$Mean <- c((mean(sample2_size10$p_hat)),(mean(sample2_size50$p_hat)),(mean(sample2_size100$p
SampleStats2$p <- c(0.20)
SampleStats2$SE <- st_error(SampleStats2$Size, SampleStats2$p)
```

```
SampleStats2 <- SampleStats2[,2:6]
SampleStats2
```

```
##   Size  Reps      Mean   p         SE
## 1   10 50000 0.2234703 0.2 0.12649111
## 2   50 50000 0.2000416 0.2 0.05656854
## 3  100 50000 0.1999776 0.2 0.04000000
## 4 1000 50000 0.2000251 0.2 0.01264911
```

Based on these new calculations, increasing the number of simulations does not change the mean or standard error values.

There are minuscule differences in the means and standard errors calculated in this new data frame but these changes are likely to have come from the new randomly selected samples. No changes are sufficiently large enough to change the mean or standard error beyond a few thousandths of a decimal.

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?

```
set.seed(09702020)
sample_ben_15 <- global_monitor %>%
  sample_n(15)
sample_ben_15 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           12   0.8
## 2 Doesn't benefit     3   0.2
```

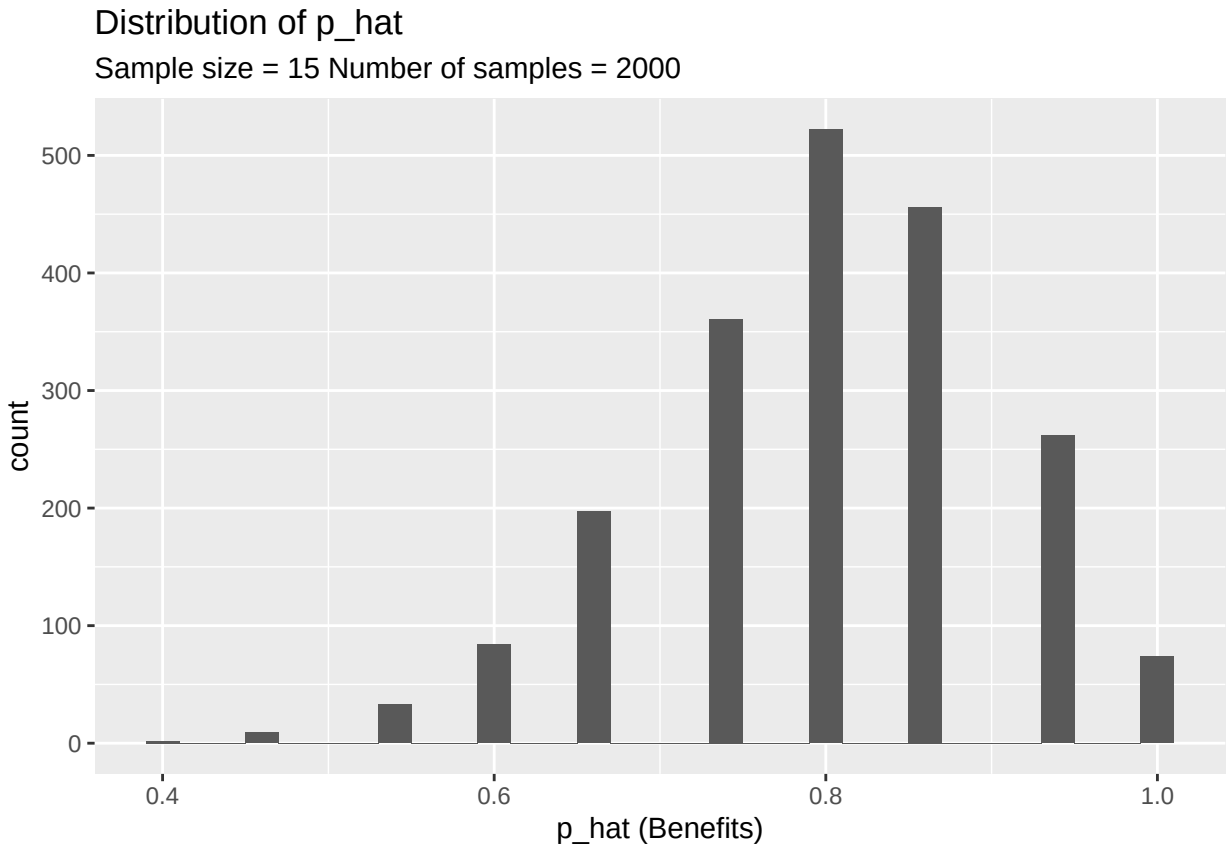
In this case, the best point estimate is based on this sample proportion at 0.80. It happens to match the population proportion but even if it did not, this would be the best point estimate because in reality we are unlikely to be able to sample the entire population.

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. 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.

```r
# At sample size 15 for 2000 samples
set.seed(09802020)
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 = "Distribution of p_hat",
    subtitle = "Sample size = 15 Number of samples = 2000"
  )
```



The shape of the data is slightly skewed with a tail extending to the left. Based on this proportion, my best guess at the population proportion would be 0.80 for those who think the work scientist do benefits them. This was selected by choosing the column with the most frequent set of values, which was right of center at 0.80. To see how close this is, we can create a proportion using the actual population.

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

```
## # A tibble: 2 x 3
```

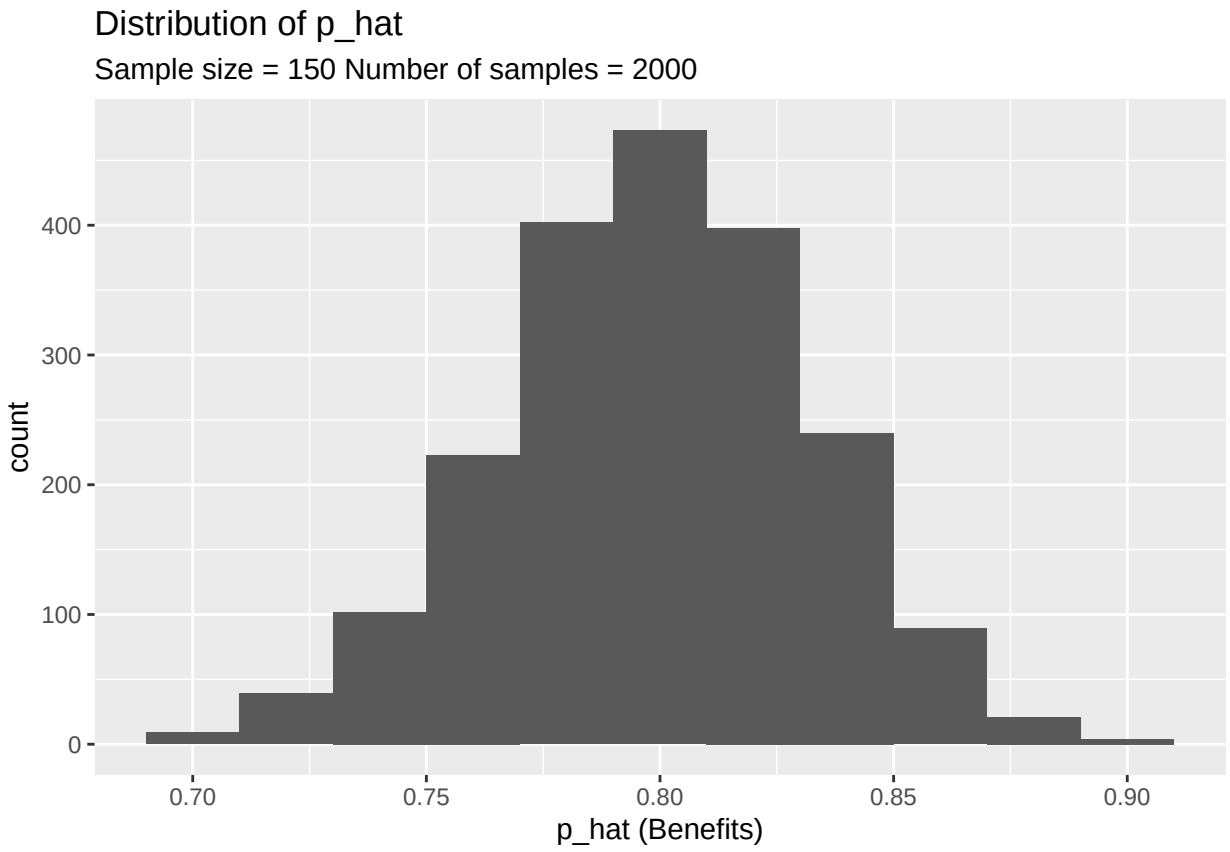
```
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits        80000   0.8
## 2 Doesn't benefit 20000   0.2
```

As it turns out, the population proportion and the estimate based on the sample proportion were identical at 0.80. This produces a sample error of 0, since, in this case, they were the same proportions.

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 enchances their lives?

```
set.seed(09812020)
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 = "Distribution of p_hat",
    subtitle = "Sample size = 150 Number of samples = 2000"
  )
```



Here again, I would estimate the population proportion to be 0.80 given this sample distribution's most frequent count towards the center of all of the most frequent counts. In this case, the distribution is symmetric and much narrower with a concentration of data around the proportion of 0.80. While the range of the sample_props15 extended from 0.40 to 1.0 the range here is smaller going from 0.70 to 0.90. Given that the population parameter remains 0.80, we would again have estimated correctly. However, with this larger sample size, our accuracy at predicting the population proportion increases.

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?

Of the distributions from 2 and 3, sample_props150 had a smaller spread than sample_props15. If I were concerned with making estimates that are more often close to the true value, I would prefer a sampling distribution with a smaller spread because there would be fewer opportunities to report proportions that are farther from the true value.

. . .