Foundations for statistical inference - Sampling distributions

Alex Ptacek

In this lab, you will investigate the ways in which the statistics from a random sample of data can serve as point estimates for population parameters. We're interested in formulating a *sampling distribution* of our estimate in order to learn about the properties of the estimate, such as its distribution.

Setting a seed: We will take some random samples and build sampling distributions in this lab, which means you should set a seed at the start of your lab. If this concept is new to you, review the lab on probability.

```
set.seed(606)
```

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages. We will also use the **infer** package for resampling.

Let's load the packages.

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

The data

A 2019 Gallup report states the following:

> The premise that scientific progress benefits people has been embodied in discoveries throughout the ages – from the development of vaccinations to the explosion of technology in the past few decades, resulting in billions of supercomputers now resting in the hands and pockets of people worldwide. Still, not everyone around the world feels science benefits them personally.
>
> **Source:** World Science Day: Is Knowledge Power?

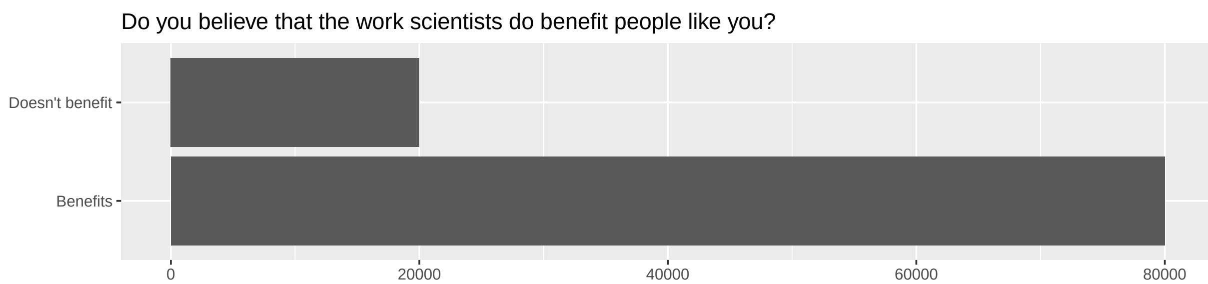
The Wellcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from the population. We will first create our population assuming a population size of 100,000. This means 20,000 (20%) of the population think the work scientists do does not benefit them personally and the remaining 80,000 think it does.

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

The name of the data frame is `global_monitor` and the name of the variable that contains responses to the question *"Do you believe that the work scientists do benefit people like you?"* is `scientist_work`.

We can quickly visualize the distribution of these responses using a bar plot.

```
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()
```



We can also obtain summary statistics to confirm we constructed the data frame correctly.

```
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
```

The unknown sampling distribution

In this lab, you have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.

If you are interested in estimating the proportion of people who don't think the work scientists do benefits them, you can use the `sample_n` command to survey the population.

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

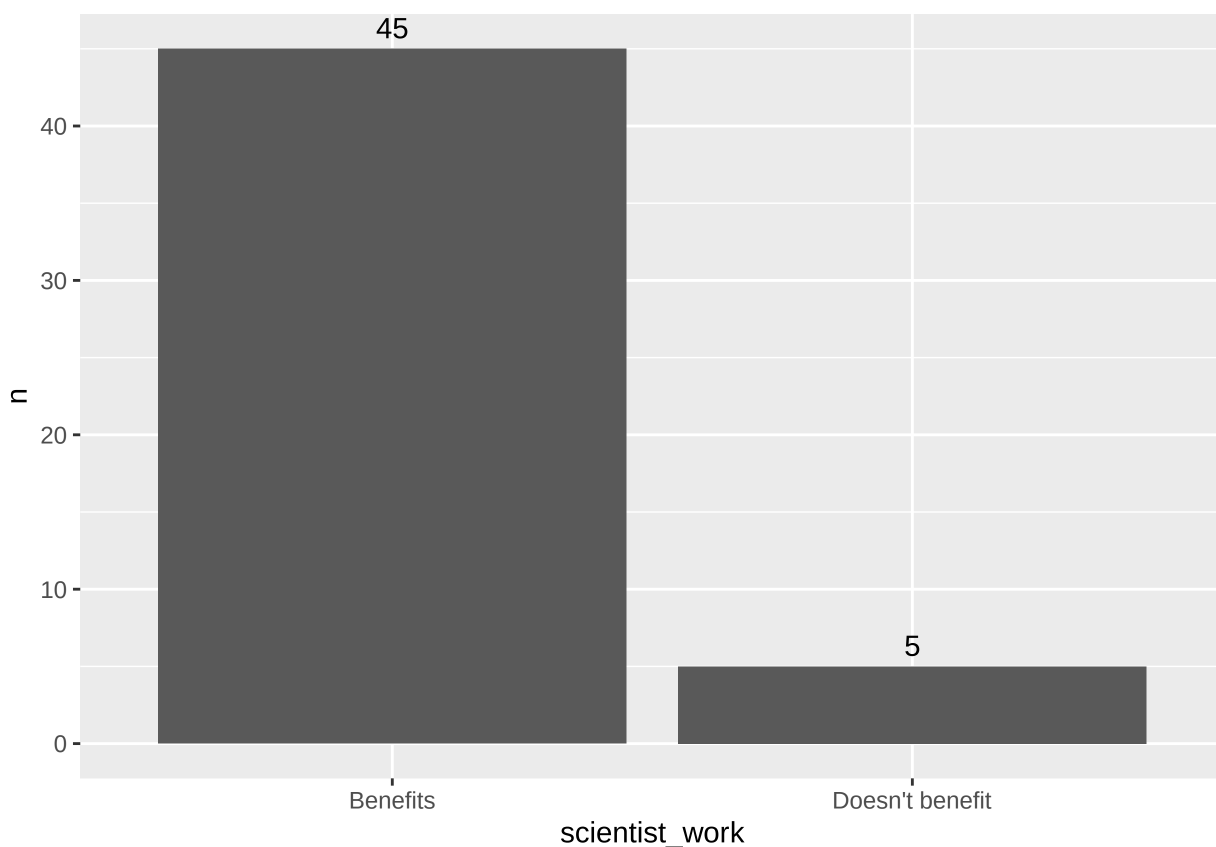
This command collects a simple random sample of size 50 from the `global_monitor` dataset, and assigns the result to `samp1`. This is similar to randomly drawing names from a hat that contains the names of all in the population. Working with these 50 names is considerably simpler than working with all 100,000 people in the population.

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.

Insert your answer here

Answer: The distribution shows that 10% of this sample thinks science benefits them versus 90% that think it does not. The 10% is 10 percentage points less than the population of 20% for the same response, which can be small or large depending on further analysis and more info.

```
samp1 |>
  count(scientist_work) |>
  ggplot(aes(x = scientist_work, y = n)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = n), vjust = -.5)
```



```
samp1 |>
  count(scientist_work) |>
  rename(n_samp = n) |>
  mutate(p_hat = n_samp/sum(n_samp))
```

```
## # A tibble: 2 x 3
##   scientist_work  n_samp p_hat
##   <chr>            <int> <dbl>
## 1 Benefits            45   0.9
## 2 Doesn't benefit      5   0.1
```

If you're interested in estimating the proportion of all people who do not believe that the work scientists do benefits them, but you do not have access to the population data, your best single guess is the sample mean.

```
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           45   0.9
## 2 Doesn't benefit     5   0.1
```

Depending on which 50 people you selected, your estimate could be a bit above or a bit below the true population proportion of 0.1. In general, though, the sample proportion turns out to be a pretty good estimate of the true population proportion, and you were able to get it by sampling less than 1% of the population.

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.

Insert your answer here

Answer: I would not expect the sample proportion to match another students, because there are a wide range of possible proportions, even though some may be highly implausible. Furthermore, I would expect another student's sample to be very different because the 95% confidence interval for my sample (1.7%, 18.3%) does not even include the population proportion of 20%. This was confirmed by another student who said their sample proportion of "Doesn't benefit" was 26%.

```
#Sample 95% Confidence Interval
ci_samp = c(.1 -  (1.96 * (sqrt(.1 * (1 - .1) / 50))), .1 +  (1.96 * (sqrt(.1 * (1 - .1) / 50)))) |>
  round(3)
ci_samp
```

```
## [1] 0.017 0.183
```

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?

Insert your answer here

Answer: The samp2 proportion is somewhat close to both the samp1 proportion and the population. The sample size of 1000 would provide a more accurate estimate of the population than sizes of 100 or 50, because samples improve in estimation accuracy of the population as they increase in size (up to a certain point, i.e. if the sample size exceeds 10% of the population, accuracy may decrease – according to the OpenIntro textbook).

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

samp2_p_hat <- samp2 |>
  count(scientist_work) |>
  rename(n_samp = n) |>
  mutate(p_hat = n_samp/sum(n_samp)) |>
  filter(scientist_work == "Doesn't benefit") |>
  print()
```

```
## # A tibble: 1 x 3
##   scientist_work  n_samp p_hat
##   <chr>            <int> <dbl>
## 1 Doesn't benefit      9  0.18
```

Not surprisingly, every time you take another random sample, you might get a different sample proportion. It's useful to get a sense of just how much variability you should expect when estimating the population mean this way. The distribution of sample proportions, called the *sampling distribution (of the proportion)*, can help you understand this variability. In this lab, because you have access to the population, you can build up the sampling distribution for the sample proportion by repeating the above steps many times. Here, we use R to take 15,000 different samples of size 50 from the population, calculate the proportion of responses in each sample, filter for only the *Doesn't benefit* responses, and store each result in a vector called `sample_props50`. Note that we specify that `replace = TRUE` since sampling distributions are constructed by sampling with replacement.

```
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")
```

And we can visualize the distribution of these proportions with a histogram.

```
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"
  )
```

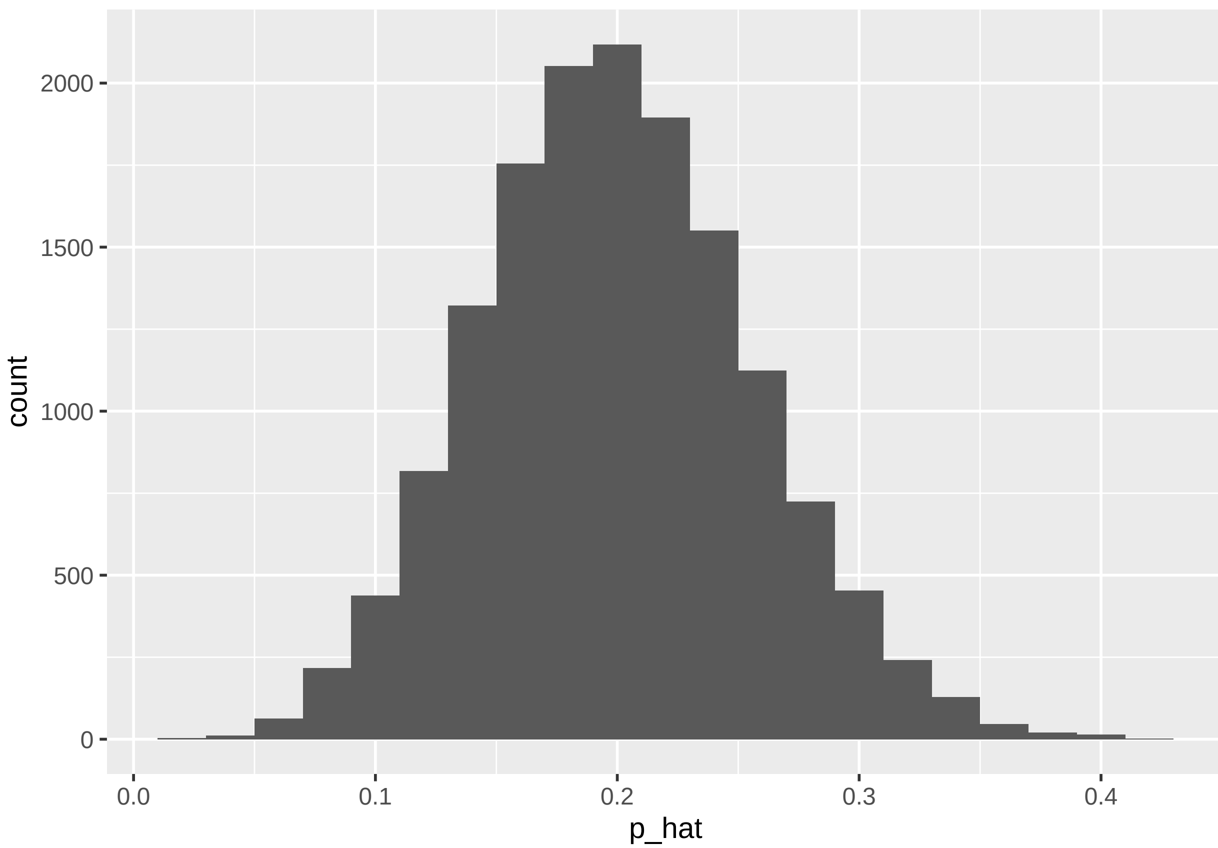
Next, you will review how this set of code works.

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.

Insert your answer here

Answer: ***sample_props50*** *has 15,000 observations and 4 variables. The sampling distribution looks nearly normal and the center is about .2, which would be the same as the population proportion.*

```
sample_props50 |>
  ggplot(aes(x = p_hat)) +
  geom_histogram(binwidth = .02)
```



Interlude: Sampling distributions

The idea behind the `rep_sample_n` function is *repetition*. Earlier, you took a single sample of size `n` (50) from the population of all people in the population. With this new function, you can repeat this sampling procedure `rep` times in order to build a distribution of a series of sample statistics, which is called the **sampling distribution**.

Note that in practice one rarely gets to build true sampling distributions, because one rarely has access to data from the entire population.

Without the `rep_sample_n` function, this would be painful. We would have to manually run the following code 15,000 times

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

```
## # A tibble: 1 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Doesn't benefit    13  0.26
```

as well as store the resulting sample proportions each time in a separate vector.

Note that for each of the 15,000 times we computed a proportion, we did so from a **different** sample!

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?

Insert your answer here

Answer: There are 23 observations in `sample_props_small`. Each observation represents one sample.

```
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")
print(sample_props_small, n = 6)
```

```
## # A tibble: 21 x 4
## # Groups:   replicate [21]
##   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     4   0.4
## 5         5 Doesn't benefit     3   0.3
## 6         7 Doesn't benefit     4   0.4
## # i 15 more rows
```

Sample size and the sampling distribution

Mechanics aside, let's return to the reason we used the `rep_sample_n` function: to compute a sampling distribution, specifically, the sampling distribution of the proportions from samples of 50 people.

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

The sampling distribution that you computed tells you much about estimating the true proportion of people who think that the work scientists do doesn't benefit them. Because the sample proportion is an unbiased estimator, the sampling distribution is centered at the true population proportion, and the spread of the distribution indicates how much variability is incurred by sampling only 50 people at a time from the population.

In the remainder of this section, you will work on getting a sense of the effect that sample size has on your sampling distribution.

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.)

Insert your answer here

Answer: Each observation in the sampling distribution represents one sample. As the sample size increases, the mean gets closer to the population, the standard error gets smaller, and the shape of the distribution looks increasingly normal. Changing the number of simulations does not affect the mean or standard error. The shape of the distribution, however, can be affected by the number of simulations. Specifically, at lower levels, the distribution doesn't have much recognizable shape. Once the the number of simulations reaches a critical point where we can easily see a nearly normal sampling distribution, increasing it further does not make it look any more normal, but the shape slightly changes.

More Practice

So far, you have only focused on estimating the proportion of those you think the work scientists doesn't benefit them. Now, you'll try to estimate the proportion of those who think it does.

Note that while you might be able to answer some of these questions using the app, you are expected to write the required code and produce the necessary plots and summary statistics. You are welcome to use the app for exploration.

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?

Insert your answer here

Answer: Based on the simulated sample below, our best estimate of the population proportion is 20%.

```r
global_monitor |>
  sample_n(15) |>
  count(scientist_work) |>
  mutate(p_hat = n/sum(n)) |>
  filter(scientist_work == "Doesn't benefit")
```

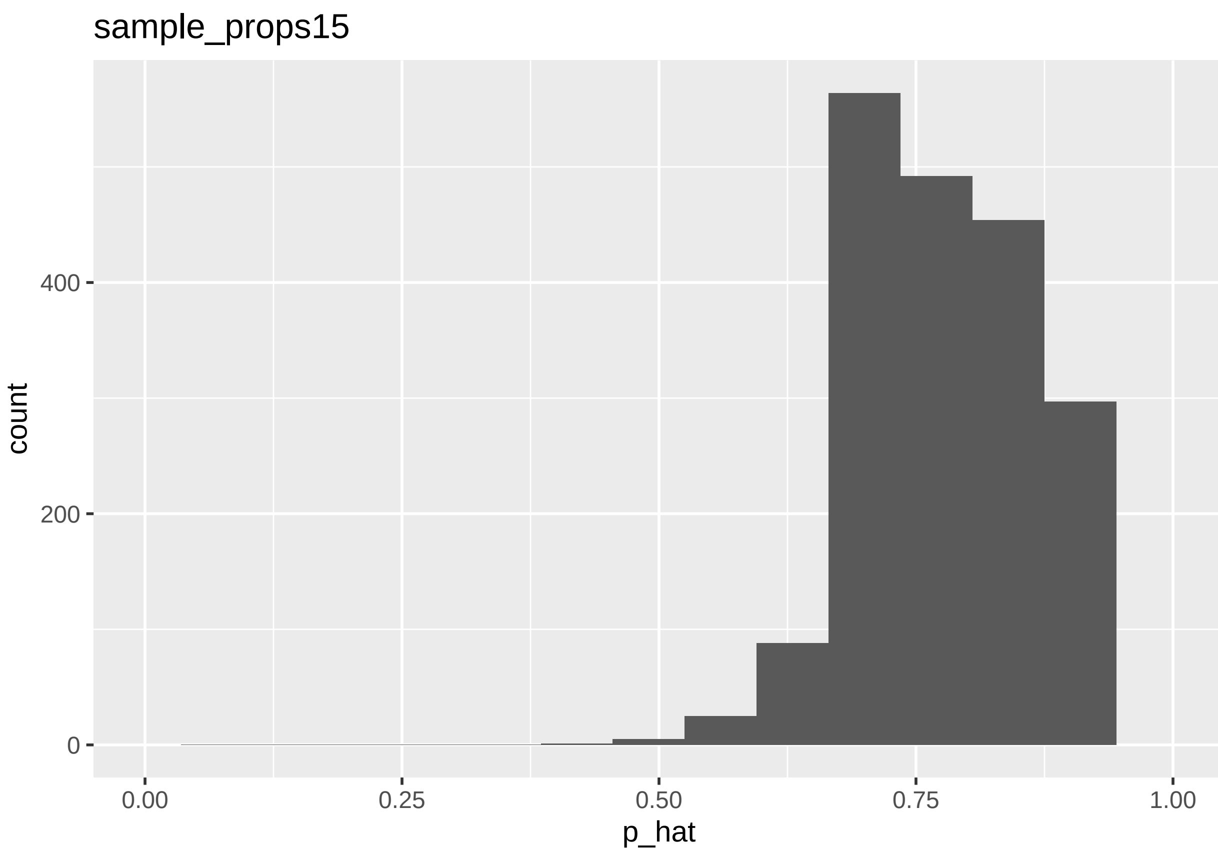
```
## # A tibble: 1 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Doesn't benefit     3   0.2
```

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.

Insert your answer here

Answer: The shape of this sampling distribution seems slightly left skewed, but somewhat normal. Based on the histogram, I would guess that the population proportion is about 80%. The sampling proportion is 80.2% which was close to my guess. The population proportion is 80%.

```
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")

sp15 <- sample_props15 |>
  ggplot(aes(x = p_hat)) +
  geom_histogram(binwidth = .07) +
  scale_x_continuous(limits = c(0, 1)) +
  labs(title = "sample_props15")
sp15
```




```
#Sample proportion (p_hat)
sample_props15 |>
  group_by(scientist_work) |>
  summarise(n = sum(n), p_hat = n/30000) |>
  select(!n)
```

```
## # A tibble: 1 x 2
```

```
##   scientist_work p_hat
##   <chr>          <dbl>
## 1 Benefits       0.803
```

```
#Population proportion (p)
global_monitor |>
  group_by(scientist_work) |>
  count() |>
  mutate(p = n/100000) |>
  filter(scientist_work == "Benefits") |>
  select(!n)
```

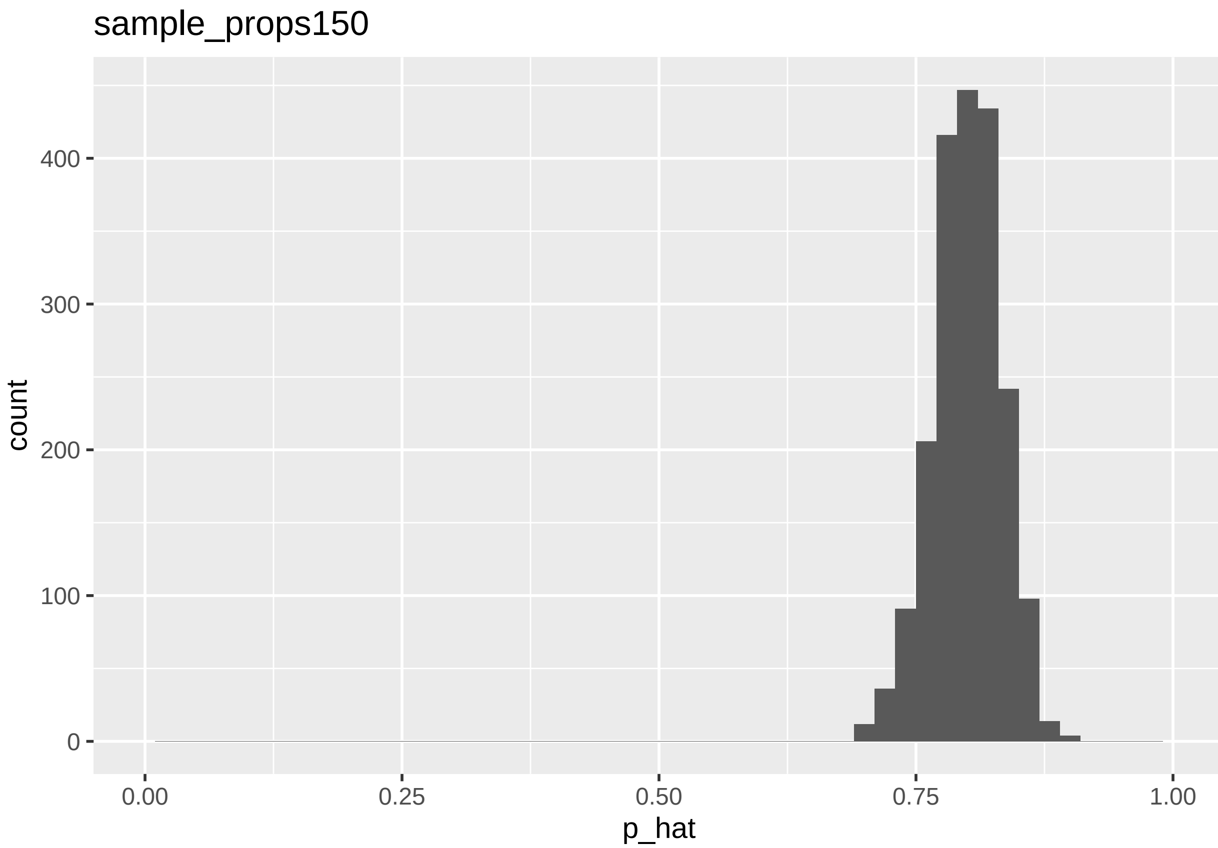
```
## # A tibble: 1 x 2
## # Groups:   scientist_work [1]
##   scientist_work     p
##   <chr>          <dbl>
## 1 Benefits         0.8
```

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?

Insert your answer here

Answer: The shape of the `sample_props150` sampling distribution looks nearly normal and seems to have a center slightly above 80%. Compared to `sample_props15`, this distribution looks much more normal, but the center looks very similar. Based on this sampling distribution, I would guess that the population proportion is about 81%.

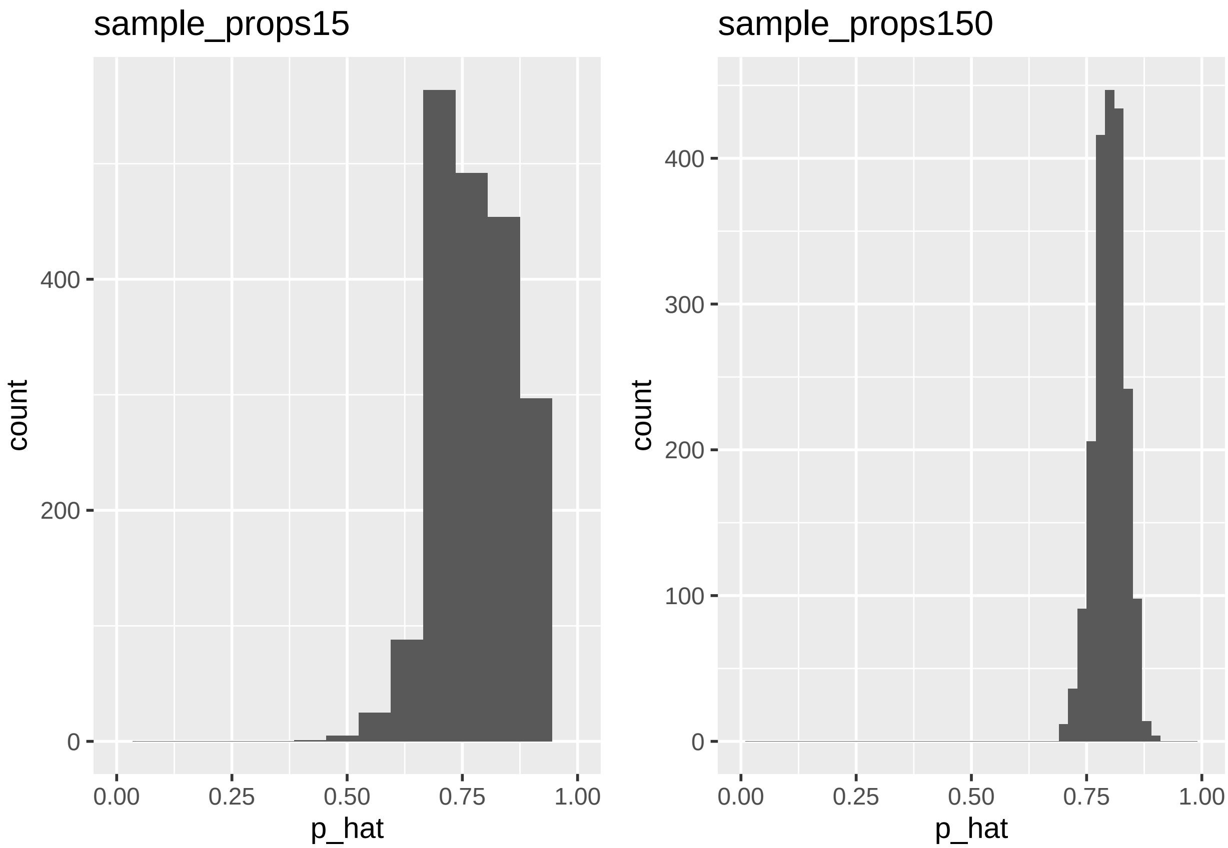
```
sample_props150 <- global_monitor |>
  rep_sample_n(size = 150, reps = 2000, replace = TRUE) |>
  count(scientist_work) |>
  mutate(p_hat = n/sum(n)) |>
  filter(scientist_work == "Benefits")

sp150 <- sample_props150 |>
  ggplot(aes(x = p_hat)) +
  geom_histogram(binwidth = .02) +
  scale_x_continuous(limits = c(0, 1)) +
  labs(title = "sample_props150")
sp150
```



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?

Answer: Between `sample_props15` and `sample_props150`, the lower sample size gives us the larger spread. If we want to make estimates that more often capture or are close to the true value, we would want a wider spread, so we can use smaller sample sizes.

```r
ggarrange(sp15, sp150)
```


sample_props15
count
400
200
0
0.00
0.25
0.50
0.75
1.00
p_hat
sample_props150
count
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
p_hat