Foundations for statistical inference - Sampling Distributions

completion time: approx 7 hours

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.

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)
```

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?

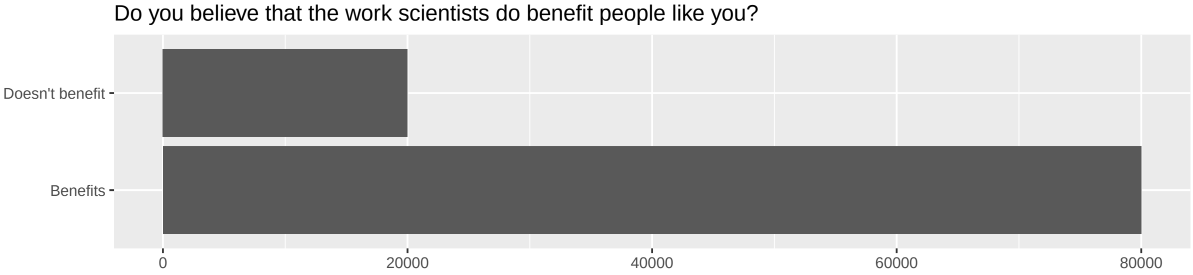
The Welcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from 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.

---

1. Insert your answer here

Any few samples you take is going to vary a bit, especially if the sample size is smaller. I think my sample is about what I would expect from a data set with an original sample percentage is 80%/20%. I ran it twice to check myself.

```
# > samp1 <- global_monitor %>% sample_n(50)
# > samp1  %>% count(scientist_work) %>% mutate(p = n /sum(n))
# # A tibble: 2 × 3
#   scientist_work      n     p
#   <chr>           <int> <dbl>
# 1 Benefits           39  0.78
# 2 Doesn't benefit    11  0.22

# > samp1 <- global_monitor %>% sample_n(50)
# > samp1  %>% count(scientist_work) %>% mutate(p = n /sum(n))
# # A tibble: 2 × 3
#   scientist_work      n     p
#   <chr>           <int> <dbl>
# 1 Benefits           43  0.86
# 2 Doesn't benefit     7  0.14
```

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           37  0.74
2 Doesn't benefit    13  0.26
```

Depending on which 50 people you selected, your estimate could be a bit above or a bit below the true population proportion of 0.26. 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.

---

2. Insert your answer here

I wouldn't expect any other samples to be the same, but I would expect student sample ratios to fall somewhere around 80/20, maybe with about the hightest variation being 15, so 65-35, or 95-5, but mostly fairly close to 80/20 maybe plus or minus 7%.

---

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?

---

3. Insert your answer here

This is a great fundamental question. It asks us to think about the differences between sample sizes and ending ratios, and how few sample points is enough to get an accurate-enough picture of the population. And, I did just ran that resampling about 100 times with sample sizes of 10, 30, 100, 1000 so I get why this question is here. The first thing I noticed is that it is really easy to see that the n=10 isn't enough. All those ratios were only on decimal point deep, and that differences jumped all the way from 100%/0% to 50%/50%. N=30 and n=100 were much closer together, though the thirty was a little more variable. At n=1000, it's really close all the time, perhaps a maximum of 2 percentage points a way from 80/20.

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.

```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")
```

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

```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"
#   )
```

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.

---

4. Insert your answer here

For the counts, it's better to understand them at every line of code

1. 750,000 - `rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%`

- Explodes the data frame to 50 * 15,000 records
- Adds a column called "replicate" which is an auto-incrementing integer for each of the 15,000 iterations.
- Think of it like, pull 15,000 groups of 50 records, where each record of the 50 is either benefit or doesn't benefit.

2. 30,000 - `count(scientist_work) %>%`

- This is a group by, and it filters it down to the 30,000 records – which is 15,000 benefit and 15,000 doesn't benefit.

3. 30,000 - `mutate(p_hat = n /sum(n)) %>%`

- Adds a column but does not change the count

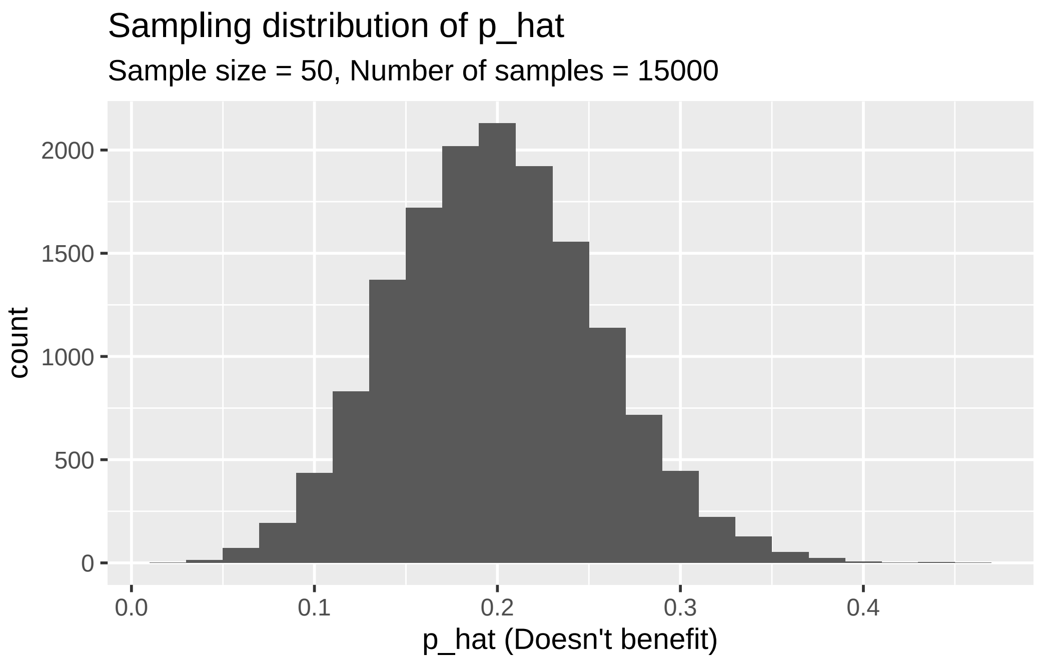
4. 15,000 - `filter(scientist_work == "Doesn't benefit")`

- removes the 15K that is Benefits.

This chart shows pretty clearly there is a normal distribution about around the 20% of the time mark.

```
# data looks like this.
# > 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    14  0.28
#  2         2 Doesn't benefit    12  0.24
#  3         3 Doesn't benefit    15  0.3
#  4         4 Doesn't benefit    15  0.3
#  5         5 Doesn't benefit    14  0.28
#  6         6 Doesn't benefit     6  0.12
#  7         7 Doesn't benefit     8  0.16
#  8         8 Doesn't benefit     8  0.16
#  9         9 Doesn't benefit     6  0.12
# 10        10 Doesn't benefit    12  0.24
```

```r
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(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"
  )
```



Sampling distribution of p_hat

Sample size = 50, Number of samples = 15000

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?

---

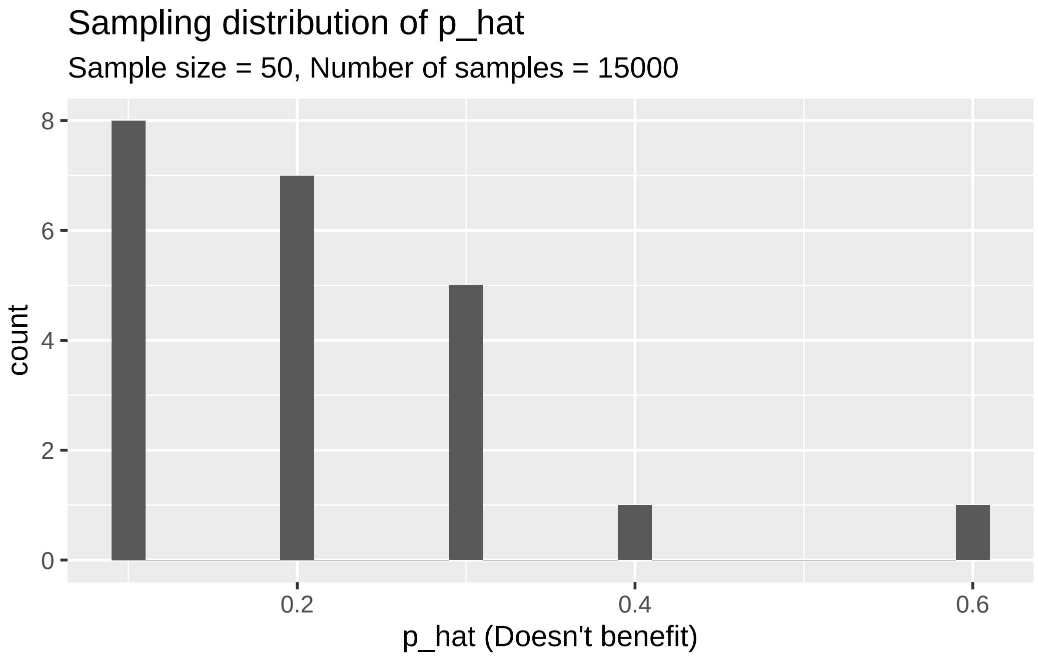
5. Insert your answer here

As before, the total size of set produced by rep_sample_n is the size(n = 10) repetitions(reps=25), so 250. It's like a combination problem, like C(Population, 5, replace=true), but done 25 times.*

Each observation represents what's in the record, test number (one of twenty five), and the percentage of times that the record "Doesn't Benefit" was in that test.

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

  ggplot(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 = 50, Number of samples = 15000"
  )
```



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.)

---

6. Insert your answer here

- What does each observation in the sampling distribution represent? *Answered in question 5 above.*
- How does the mean, standard error, and shape of the sampling distribution change as the sample size increases? *If the sample increases, the distribution curve gets smoother and tighter around 20%.*
- 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.) *The purpose of this exercise is to run multiple simulations and see them plotted together. The more you have, the higher the counts get on the y axis, and the more bars you see filling out the space under the bell curve.*

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?

7. Insert your answer here

On this run, 10 out of 15, or 67% of believed scientists benefit their lives.

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

# A tibble: 1 × 3
  scientist_work     n p_hat
  <chr>          <int> <dbl>
1 Benefits          10 0.667
```

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.

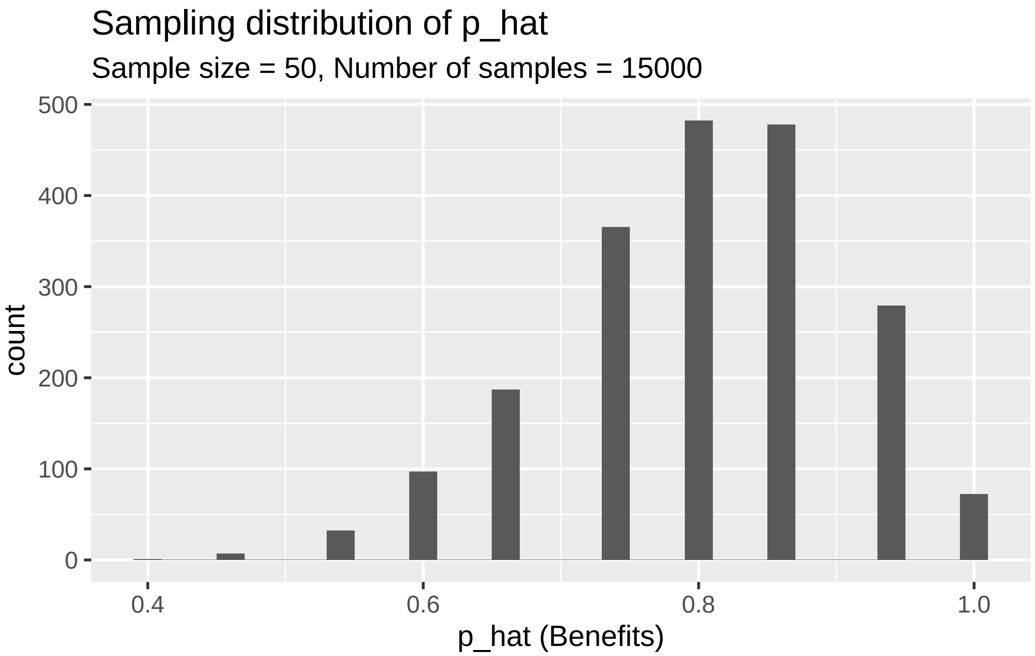
8. Insert your answer here

As discussed in question 3, this question implies that if we have a smaller sample size. Also, though, if we sample from it many times, eventually we start to find the middle anyway. We can either have big sample sizes, or we can run smaller sample sizes many times. The difference between these two tactics is dispersion.

After running this dozens of times, this code does show the mean of benefits also being around 80%, but the curve is wider and less accurate in terms of variance.

```
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(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 = 50, Number of samples = 15000"
  )
```



As to the final question, "calculate and report the population proportion" I think you're asking me to report 80%, as that's what it is in the actual population. If you're asking about the sample I just ran for this question, it was

```
> mean(sample_props15$p_hat)
[1] 0.8009667
```

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?

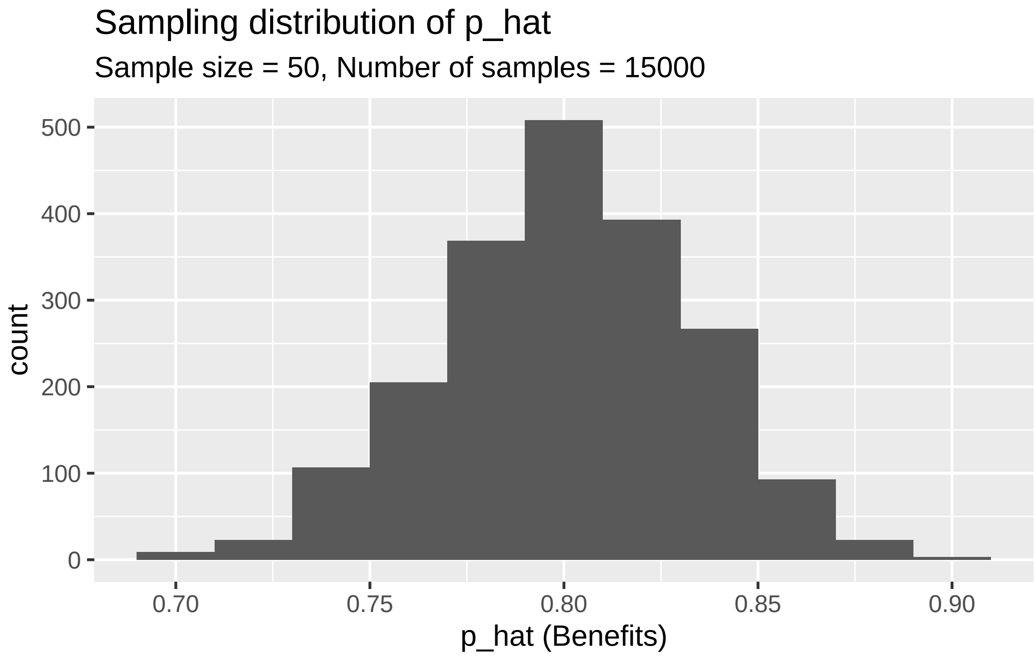
9. Insert your answer here

Visually, it again looks about like 80%, but with considerably less less variance. Just by eyeballing, I'd say it has as standard deviation of 8%, while the one with n=15 is more like

15%.

```
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(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 = 50, Number of samples = 15000"
  )
```



```
> mean(sample_props150$p_hat)
[1] 0.8004367
```

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?

Assuming you mean 8 and 9, and not 2 or 3, #9 has the smaller spread. And if spread matters, I'm going to prefer to either A, have a larger sample or B, take more samples.