Foundations for statistical inference - Sampling distributions

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?

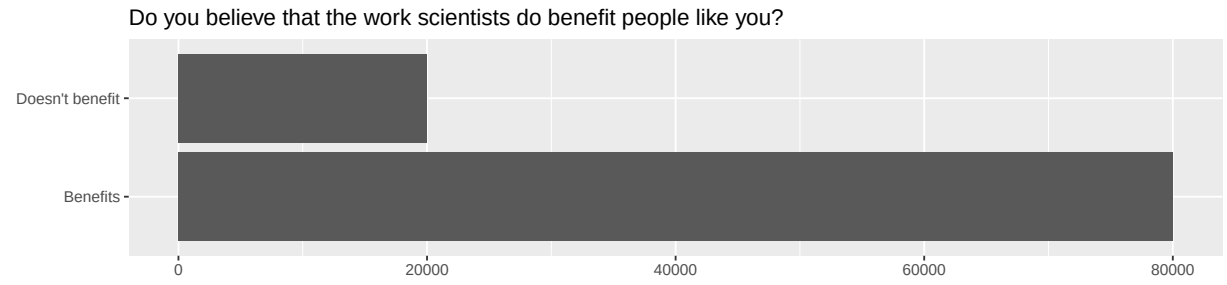
The Wellcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from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

```
set.seed(1234)
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

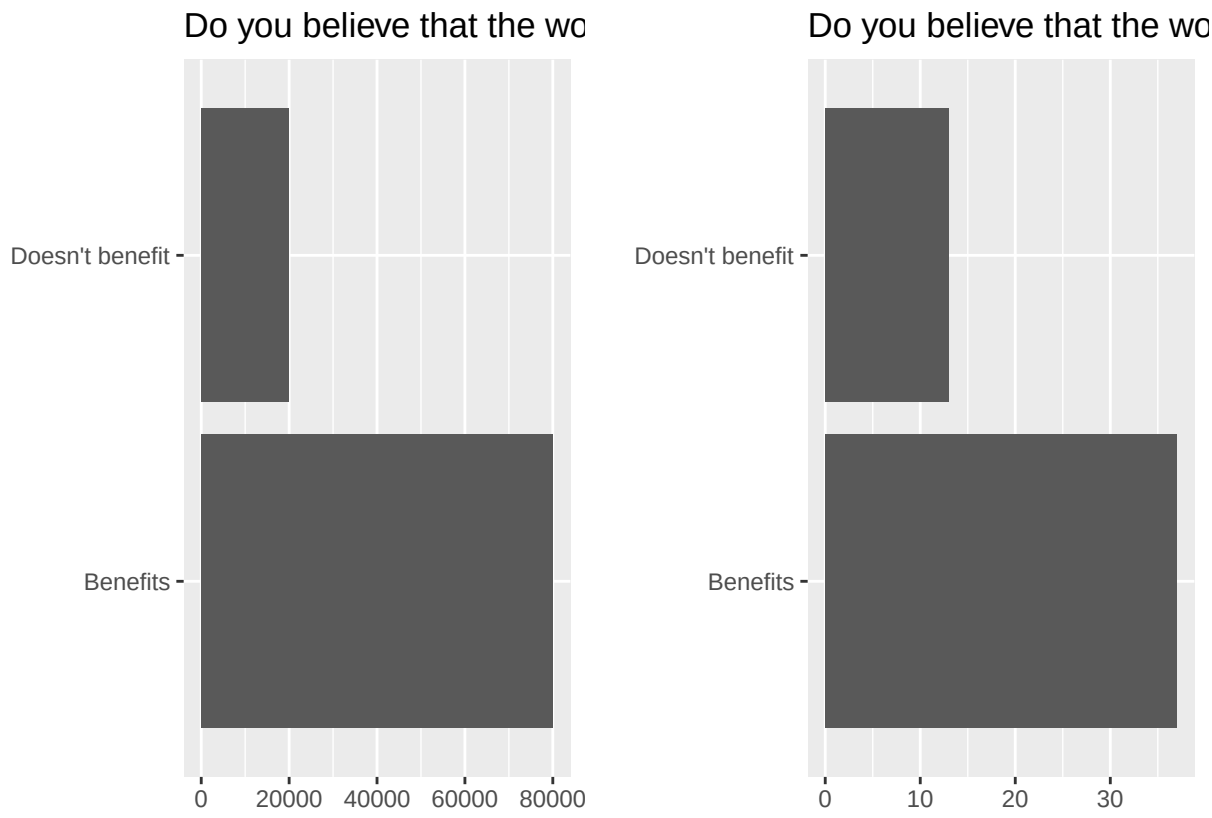
names. Code you presented earlier for visualizing and summarizing the population data will still be useful for the sample, however be careful to not label your proportion `p` since you're now calculating a sample statistic, not a population parameters. You can customize the label of the statistics to indicate that it comes from the sample.

ANSWER: - In the below cowplot I compare the distribution of responses in the sample to that of the population. The distributions are similar: A much higher proportion, roughly 3x or more, believe that the work scientists do benefits them. - They are distributed in a binomial distribution, since the two options for the survey responses were categorical: "Doesn't Benefit" or "Benefits."

```
pop_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()
```

```
sample_plot <- 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()
```

```
cowplot::plot_grid(pop_plot,sample_plot)
```



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.

ANSWER: - I'd expect my sample proportion to be similar to another student's because they are samples from the same population. Though because they are random samples (sample_n), they are likely not exactly the same distributions. - A fellow student had a sample of 0.14 while I had a sample of 0.26: They are a similar deviation away from the expected mean value of 0.20.

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?

ANSWER: - samp2's proportion is 0.24, similar to samp1's proportion of 0.26 and my classmates 0.14. - The larger the random sample, the more likely the random sample would be to represent the population its drawn from. Therefore of these samples, a sample of 1000 would likely provide the most accurate estimate of the population proportion.

```
set.seed(3)
samp2 <- global_monitor %>%
  sample_n(50)

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

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

```
# For use inline below
samp2_p_hat <- samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat2 = n /sum(n)) %>%
  filter(scientist_work == "Doesn't benefit") %>%
  pull(p_hat2) %>%
  round(2)

samp2_p_hat
```

```
## [1] 0.24
```

```
Not surprisingly, every time you take another random sample, you might get a different sample proportio
```

```
```r
set.seed(4)
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.

ANSWER: - There are 15,000 elements or observations within sample_props50, one for each sample taken. - The distribution appears to be normally distributed. It has a nearly identical mean and median of 0.199 and 0.2 respectively (calculated below in mean-median R chunk). - Plot of distribution below.

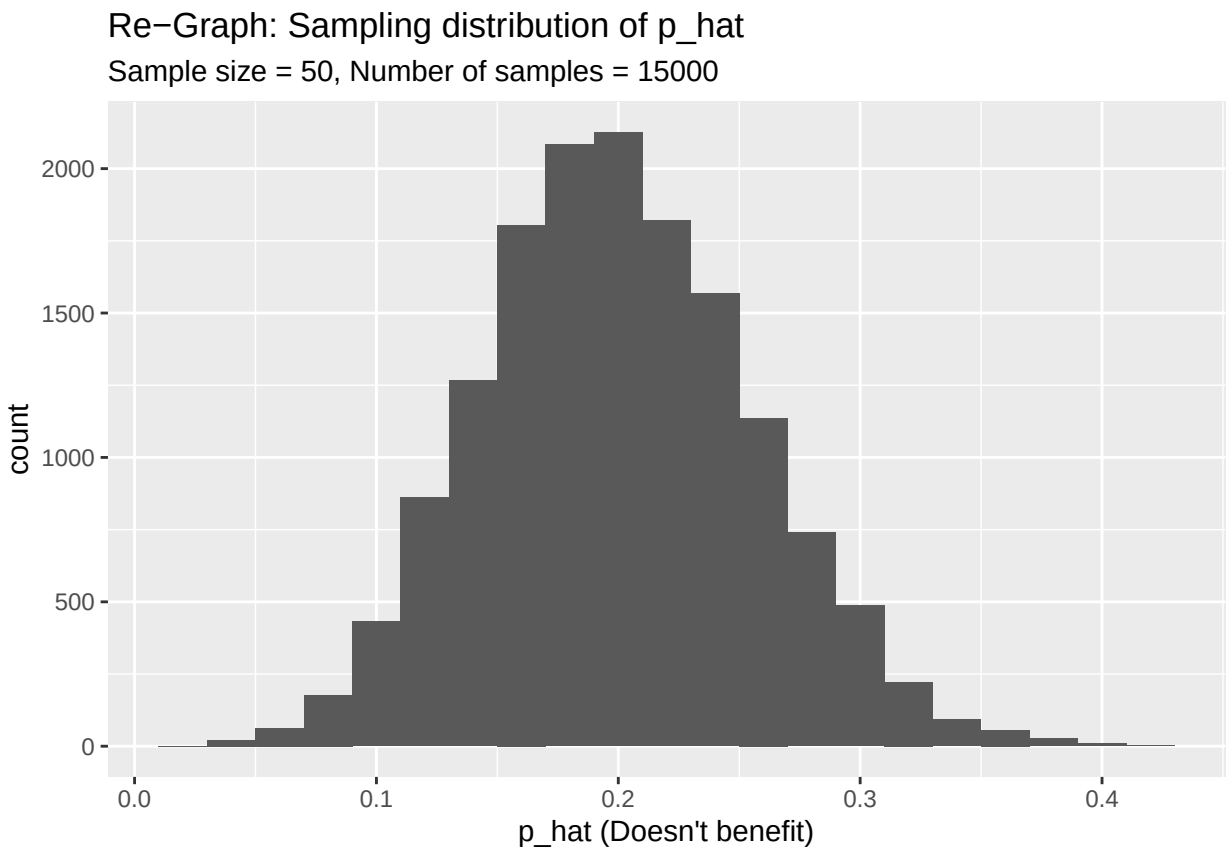
```
median(sample_props50$p_hat)
```

```
## [1] 0.2
```

```
mean(sample_props50$p_hat)
```

```
## [1] 0.1997013
```

```
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Re-Graph: Sampling distribution of p_hat",
    subtitle = "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    10   0.2
```

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?

ANSWER: - Please see below for code. - There are 25 observations in this dataframe "sample_props_small". Each observation represents a sample from the global_monitor dataframe, where 10 objects from global_monitor are taken and collectively assessed for their ratio of "Doesn't benefit" vs. "Benefits".

```
set.seed(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")

print(sample_props_small)
```

```
## # A tibble: 18 x 4
## # Groups:   replicate [18]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         2 Doesn't benefit     3   0.3
##  2         4 Doesn't benefit     2   0.2
##  3         5 Doesn't benefit     1   0.1
##  4         6 Doesn't benefit     1   0.1
##  5         8 Doesn't benefit     1   0.1
##  6        10 Doesn't benefit     1   0.1
##  7        11 Doesn't benefit     1   0.1
##  8        13 Doesn't benefit     2   0.2
##  9        14 Doesn't benefit     3   0.3
## 10        15 Doesn't benefit     4   0.4
## 11        17 Doesn't benefit     4   0.4
## 12        18 Doesn't benefit     3   0.3
## 13        20 Doesn't benefit     3   0.3
## 14        21 Doesn't benefit     1   0.1
## 15        22 Doesn't benefit     3   0.3
## 16        23 Doesn't benefit     5   0.5
## 17        24 Doesn't benefit     1   0.1
## 18        25 Doesn't benefit     1   0.1
```

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.)

ANSWER: - Each observation in the sampling distribution represents a sample from the global_monitor dataframe, where 10, 50, or 100 objects from global_monitor are taken and collectively assessed for their ratio of "Doesn't benefit" vs. "Benefits". This sampling is done 5,000 times (5,000 simulations). - The graphed "p_hat" is the proportion of people who do *not* think that scientists benefit their lives. - As the sample size increases, the mean, standard, error and shape increasingly resemble that of a normal distribution. This aligns with the Central Limit Theorem (CLT) discussed in class. CLT usually requires a sample size of greater than 30, which is also the case here as a sample size of 10 creates a graph that's too right or positive skewed. - As the number of samples increases, the graph's attributes also increasingly resemble that of a normal distribution.

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 enchances their lives?

ANSWER: - R chunk below - Using this sample, the best estimate of the population proportion who think scientists enhance their lives is 80%. I derived this from the mean below, combined with the prior samples (in R chunks above) confirming a mean of 20% do not think scientists' work benefits them.

```
set.seed(7)

sample_props_15_benefits <- global_monitor %>%
                    rep_sample_n(size = 15, reps = 25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")

sample_props_15_benefits
```

```
## # A tibble: 25 x 4
## # Groups:   replicate [25]
##    replicate scientist_work     n p_hat
##        <int> <chr>          <int> <dbl>
##  1         1 Benefits           9 0.6
##  2         2 Benefits          10 0.667
##  3         3 Benefits          11 0.733
##  4         4 Benefits          12 0.8
##  5         5 Benefits          10 0.667
##  6         6 Benefits          10 0.667
##  7         7 Benefits          12 0.8
##  8         8 Benefits          13 0.867
##  9         9 Benefits          10 0.667
## 10        10 Benefits           9 0.6
## # ... with 15 more rows
```

```
mean(sample_props_15_benefits$p_hat)
```

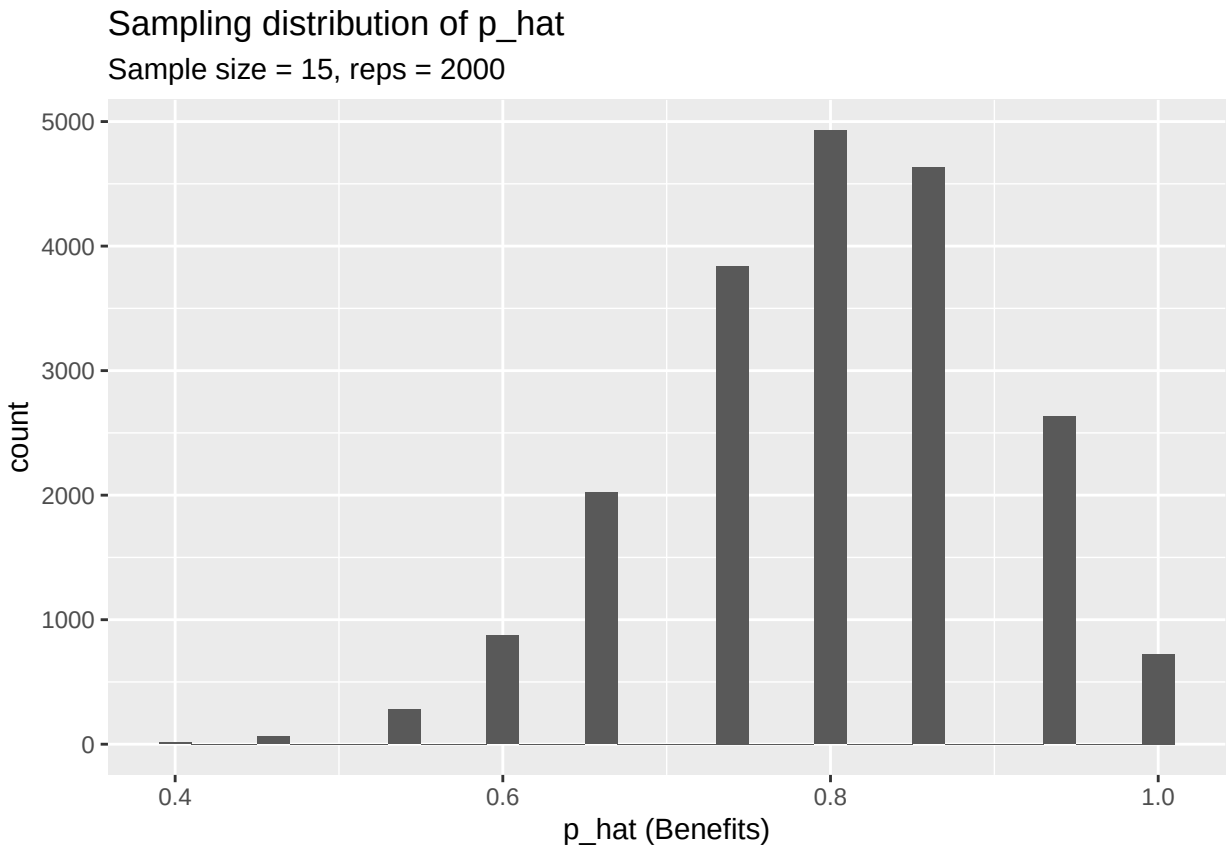
```
## [1] 0.7706667
```

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.

ANSWER: - R chunk below - The distribution is generally normal with a left or negative skew - The "true" proportion of those who think the work scientists do enhances their lives is approximately 80%. The mode, median, and mean are all this value. (Calculated below.) - The population of this survey is "global" (per welcome Global Monitor). Therefore "true population" calculation = 8 billion (global population) * 80% = 6.4B people in the world believe the work scientists do benefits them.

```
set.seed(8)
sample_props15 <- global_monitor %>%
                    rep_sample_n(size = 15, reps = 20000, 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, reps = 2000"
  )
```




```
#Mode apparent from graph: Value with highest count is p_hat of 0.8
mean(sample_props15$p_hat)
```

```
## [1] 0.8001767
```

```
median(sample_props15$p_hat)
```

```
## [1] 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 enchances their lives?

ANSWER: - R chunk below - The distribution is more normal than the sample size 15 graph, with no obvious skew, and with a taper from the mean in line with normal distributions' standard deviations. - Based on this graph, I'd be even more confident about the "true" proportion of those who think the work scientist do enhances their lives. This proportion is 80% of the world's population, or 6.4B people.

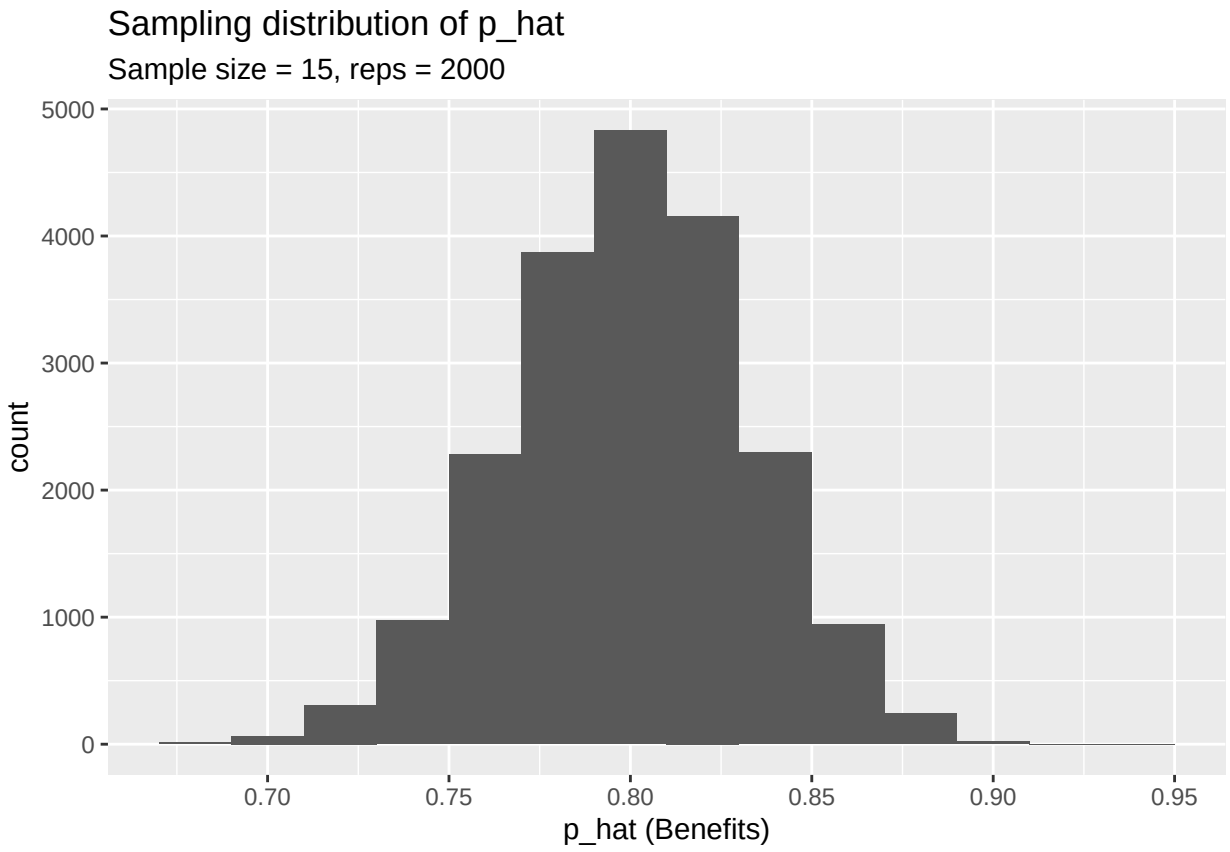
```
set.seed(9)
sample_props150 <- global_monitor %>%
                    rep_sample_n(size = 150, reps = 20000, replace = TRUE) %>%
                    count(scientist_work) %>%
```

```
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Benefits")

sample_props150
```

```
## # A tibble: 20,000 x 4
## # Groups:   replicate [20,000]
##    replicate scientist_work     n p_hat
##        <int> <chr>          <int> <dbl>
##  1         1 Benefits         125 0.833
##  2         2 Benefits         117 0.78
##  3         3 Benefits         128 0.853
##  4         4 Benefits         120 0.8
##  5         5 Benefits         120 0.8
##  6         6 Benefits         126 0.84
##  7         7 Benefits         117 0.78
##  8         8 Benefits         127 0.847
##  9         9 Benefits         126 0.84
## 10        10 Benefits         108 0.72
## # ... with 19,990 more rows
```

```
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 = 15, reps = 2000"
  )
```




```
#Mode apparent from graph: Value with highest count is p_hat of 0.8
mean(sample_props150$p_hat)
```

```
## [1] 0.7996923
```

```
median(sample_props150$p_hat)
```

```
## [1] 0.8
```

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: - Sampling distributions with smaller spreads are better for making estimates because they have a smaller range of probable values. Therefore we can be more confident in the true value. - I admit I'm not sure what's referenced in terms of the question's "sampling distributions from 2 and 3". In questions 9 and 10, the sample_props150 had a tighter distribution and smaller range. sample_props150's p_hat summary stats were: 0.67 min, 0.78 1st quartile, 0.80 median, 0.82 3rd quartile, 0.93 max compared to sample_prop15's 0.40 min, 0.73 1st quartile, 0.80 median, 0.87 3rd quartile, and 1.0 max.

```
summary(sample_props15$p_hat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.7333  0.8000  0.8002  0.8667  1.0000
```

```
summary(sample_props150$p_hat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6733  0.7800  0.8000  0.7997  0.8200  0.9333
```
