Foundations for statistical inference - Sampling distributions

Title: CUNY SPS MDS DATA606_LAB5A"

Author: Charles Ugiagbe

Date: "10/9/2021"

Load packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.1

## Warning: package 'tibble' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1
```

```
library(openintro)
```

```
## Warning: package 'openintro' was built under R version 4.1.1
```

```
library(infer)
```

```
## Warning: package 'infer' was built under R version 4.1.1
```

The data

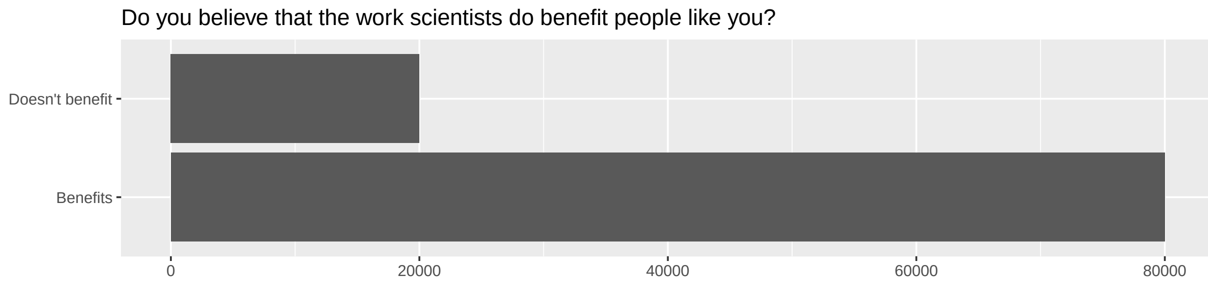
A 2019 Gallup report states the following:

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))
)
```

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

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.

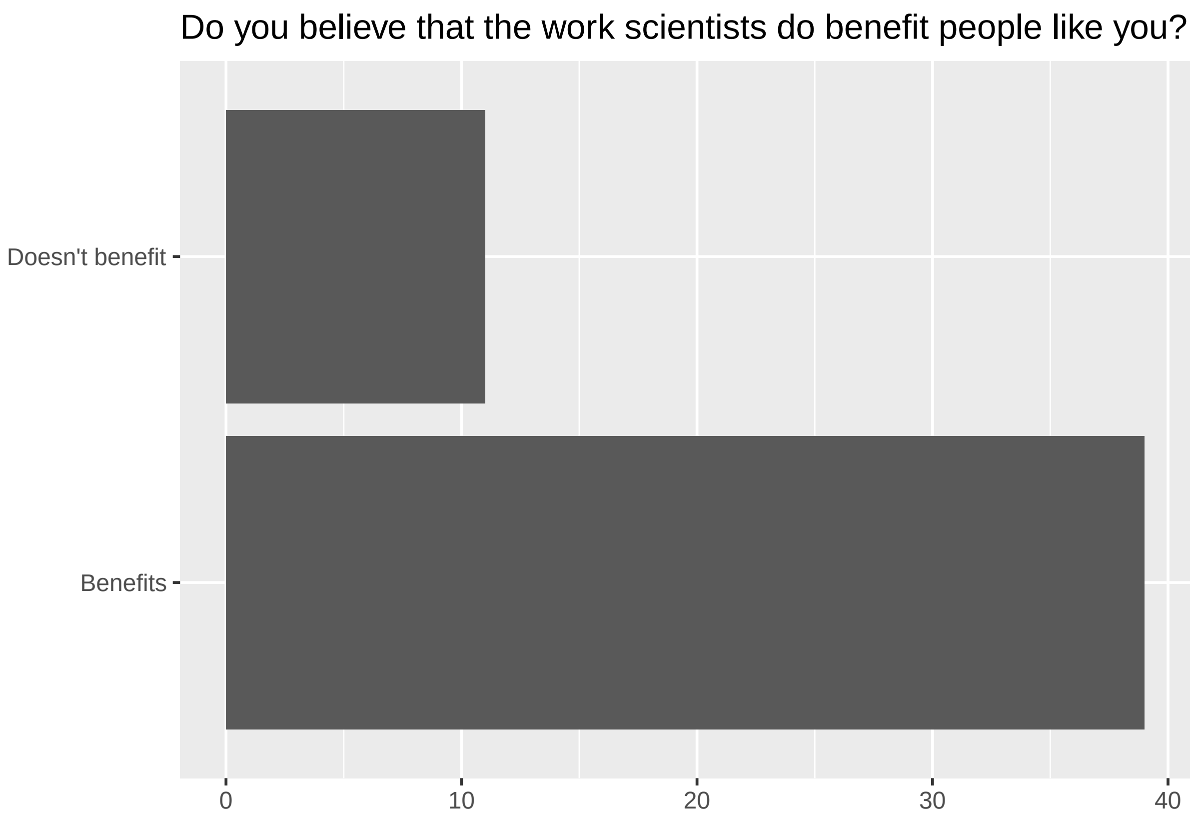
```
samp1 <- global_monitor %>%
  sample_n(50)
```

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.

Solution 1:

```
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()
```



Summary statistics for the sample data

```
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           39  0.78
## 2 Doesn't benefit    11  0.22
```

Depending on which 50 people you selected, your estimate could be a bit above or a bit below the true population proportion of 0.22. 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.

Solution 2:

I would not expect the sample proportion to match that for another students because the sample because the sample are different and ramdomly selected. The proportion would be somewhat different but similar. This is confirm from other answers.

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?

Solution 3:

Taking another sample of size 50 named "samp2"

```
set.seed(295)
samp2 <- global_monitor %>%
  sample_n(50)

samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat2 = n /sum(n))
```

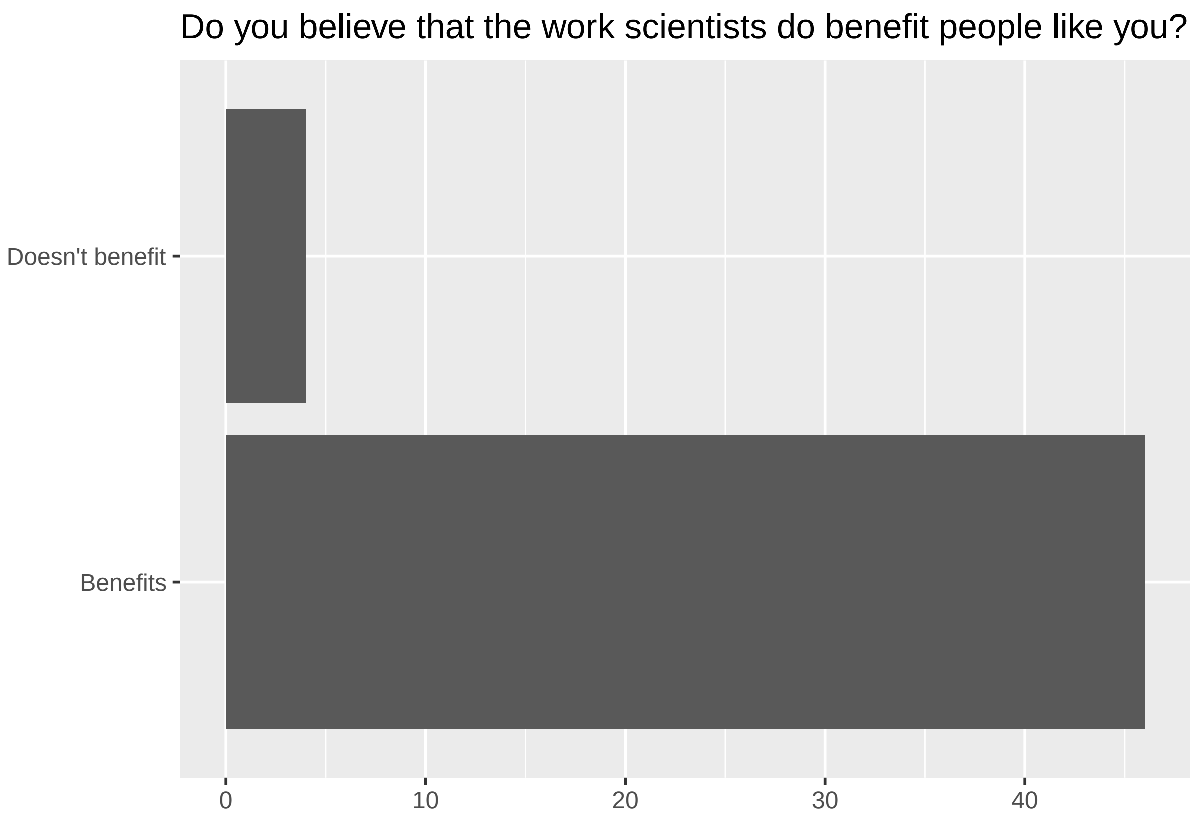
```
## # A tibble: 2 x 3
##   scientist_work      n p_hat2
##   <chr>           <int>  <dbl>
## 1 Benefits           46   0.92
## 2 Doesn't benefit     4   0.08
```

```
# 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] 0.08
```

```
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()
```



The 0.78 proportion of Benefits for sample 1 quite different from the 0.92 proportion of Benefit of sample 2. There is a decrease in the proportion of sample 1 when the second sample was examined.

If a take another two more different samples of 100 and 1000, the 1000 will produce a more accurate result because the larger the sample, the more accurate our result tend closer to the population parameter.

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.

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.

Solution 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")
```

```
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"
  )
```

```
summary(sample_props50)
```

```
##    replicate     scientist_work           n               p_hat       
##  Min.   :    1   Length:15000       Min.   : 1.000   Min.   :0.0200  
##  1st Qu.: 3751   Class :character   1st Qu.: 8.000   1st Qu.:0.1600  
##  Median : 7500   Mode  :character   Median :10.000   Median :0.2000  
##  Mean   : 7500                      Mean   : 9.997   Mean   :0.1999  
##  3rd Qu.:11250                      3rd Qu.:12.000   3rd Qu.:0.2400  
##  Max.   :15000                      Max.   :22.000   Max.   :0.4400  
```

From the summary statistics of the sampling distribution, there are 15000 elements with a mean proportion, p_hat(Doesn't benefit) of 0.2002. The mean of p_hat(Benefit) = 1 - p_hat(Doesn't benefit) = 0.7998 which is very close to the mean of the population proportion. Also,From the histogram plot, we can see that the sample is symmmentric and follow a normal distribution.

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     7  0.14
```

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?

Solution 5:

```
set.seed(2955)
sample_props_small <- global_monitor %>%
                    rep_sample_n(size = 10, reps =25, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
sample_props_small
```

```
## # A tibble: 23 x 4
## # Groups:   replicate [23]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     1   0.1
##  2         2 Doesn't benefit     2   0.2
##  3         3 Doesn't benefit     1   0.1
##  4         4 Doesn't benefit     1   0.1
##  5         5 Doesn't benefit     2   0.2
##  6         6 Doesn't benefit     1   0.1
##  7         8 Doesn't benefit     5   0.5
##  8         9 Doesn't benefit     3   0.3
##  9        10 Doesn't benefit     2   0.2
## 10        11 Doesn't benefit     4   0.4
## # ... with 13 more rows
```

There are 23 observation in the distribution and each observation represent each sample

Sample size and the sampling distribution

```
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.)

Solution 6:

For 5,000 number of Samples(simulations)

When Sample size = 10: Mean = 0.22, SE = 0.11; When Sample size = 50: Mean = 0.2, SE = 0.06; When Sample size = 100: Mean = 0.2, SE = 0.04;

As the sample size increase from 10 to 50, the mean tends to 0.2 which is the Population mean; the standard error decrease and the shape of the sampling distribution becomes more symmentric ie tends to normal.

As i increase the number of simulations, the mean and the standard error remains the same. This is so because the number of simulations is already large enough to converge the sample mean and standard error to that of the population

---

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?

Solution 7:

For sample n = 15

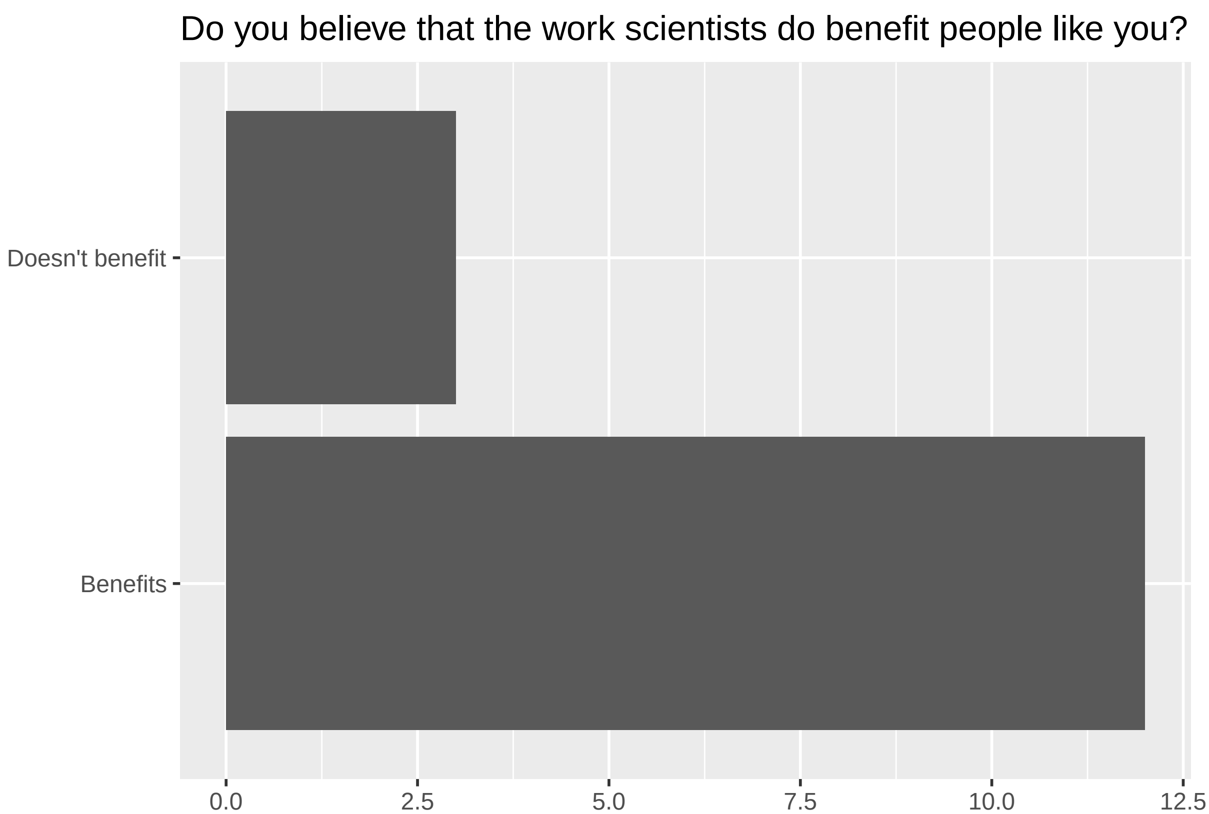
```
set.seed(4321)
samp_B15 <- global_monitor %>%
  sample_n(15)

samp_B15 %>%
  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
```

To see the Behaviour of the plot

```
ggplot(samp_B15, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```



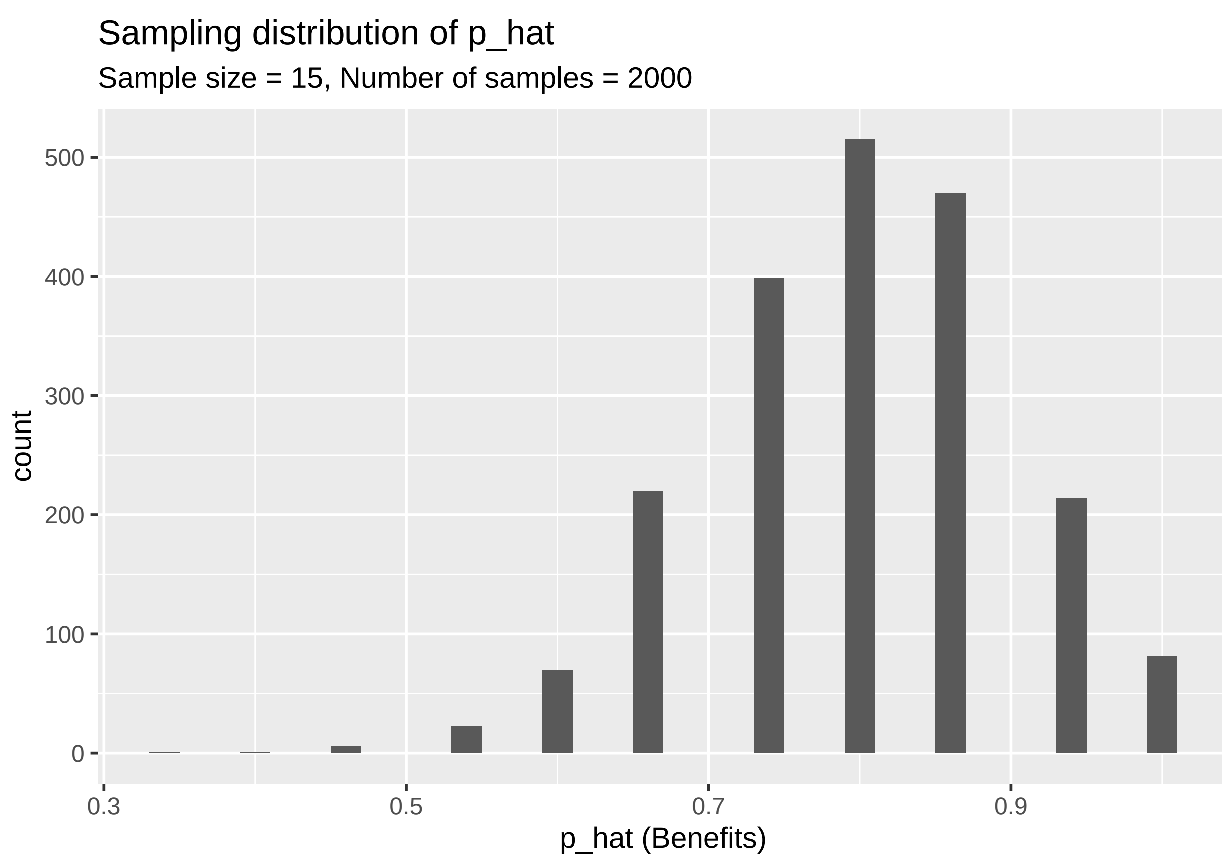
8. Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enchances their lives for samples of size 15 by taking 2000 samples from the population of size 15 and computing 2000 sample proportions. Store these proportions in as `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.

Solution 8:

```
set.seed(321)
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")
```

To see the behaviour of the Plot

```
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, Number of samples = 2000"
  )
```



check the summary of the sample

```
summary(sample_props15)
```

```
##     replicate       scientist_work             n              p_hat
```

```
##  Min.   :   1.0   Length:2000        Min.   : 5.00   Min.   :0.3333
##  1st Qu.: 500.8   Class :character   1st Qu.:11.00   1st Qu.:0.7333
##  Median :1000.5   Mode  :character   Median :12.00   Median :0.8000
##  Mean   :1000.5                      Mean   :11.98   Mean   :0.7986
##  3rd Qu.:1500.2                      3rd Qu.:13.00   3rd Qu.:0.8667
##  Max.   :2000.0                      Max.   :15.00   Max.   :1.0000
```

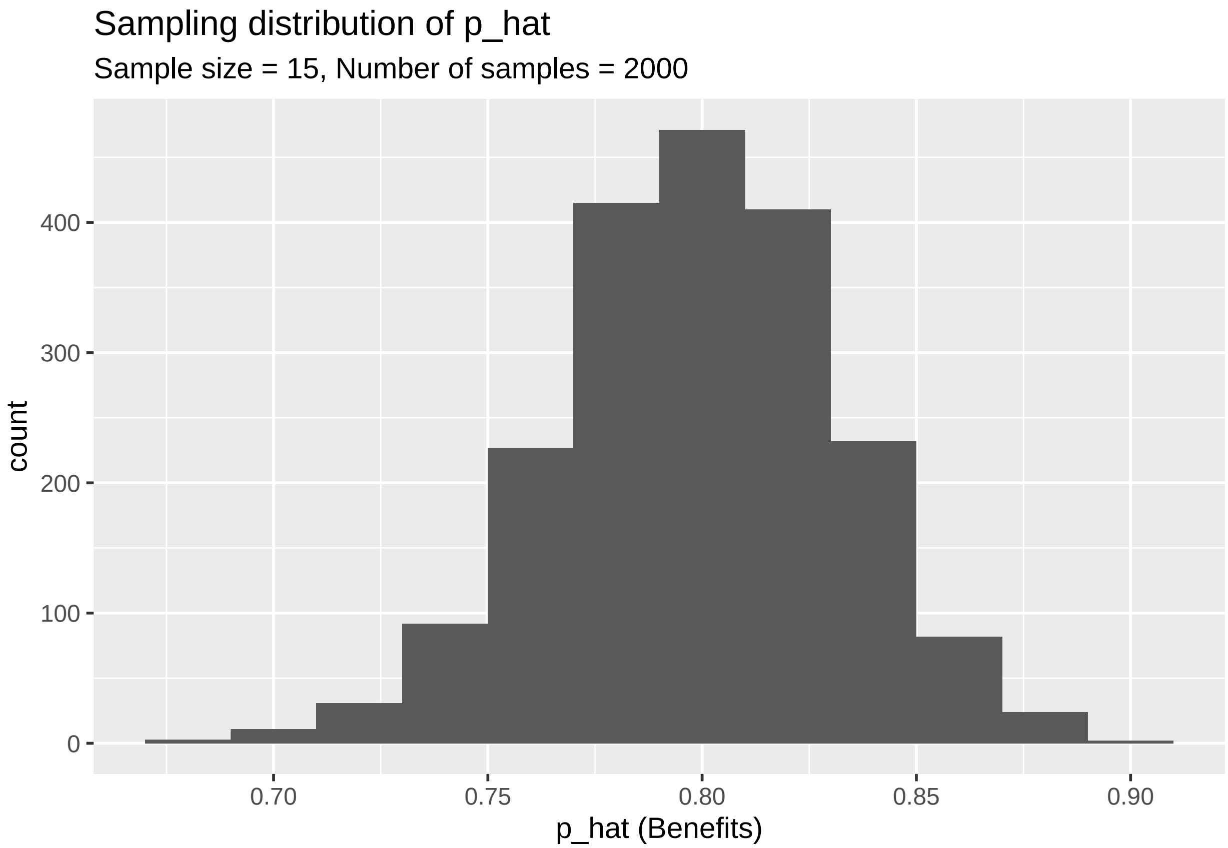
9. Change your sample size from 15 to 150, then compute the sampling distribution using the same method as above, and store these proportions in a new object called `sample_props150`. Describe the shape of this sampling distribution and compare it to the sampling distribution for a sample size of 15. 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?

Solution 9:

```
set.seed(931)
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")
```

Visualizing the Plot

```
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, Number of samples = 2000"
  )
```



check the summary of the sample

```
summary(sample_props150)
```

```
##    replicate      scientist_work           n              p_hat
##  Min.   :   1.0   Length:2000        Min.   :102.0   Min.   :0.6800
##  1st Qu.: 500.8   Class :character   1st Qu.:117.0   1st Qu.:0.7800
##  Median :1000.5   Mode  :character   Median :120.0   Median :0.8000
##  Mean   :1000.5                      Mean   :119.8   Mean   :0.7989
##  3rd Qu.:1500.2                      3rd Qu.:123.0   3rd Qu.:0.8200
##  Max.   :2000.0                      Max.   :135.0   Max.   :0.9000
```

The shape of the sampling distribution for sample size 15 look like a bar plot while that of distribution with sample size 150 look more symmmentric and tends toward a normal plot. I would guess the true population proportion to be 0.8. because the calculations of the population proportion of people who think the work scientists do enhances their lives is 0.7986 for sample size 15 and 0.7989 for sample size 150. This two value are similar and very close to 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?

Solution 10:

The sample distribution of 3 has a smaller spread. The sample Distribution with the larger sample size has the smaller spread. If i am concerned with making estimates that are more often close to true value, i would favor a sampling distribution with a large sample size and small spread