CUNY SPS DATA606 Lab5a: Foundations for statistical inference - Sampling distributions

Chinedu Onyeka

September 29th, 2021

Getting Started

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?

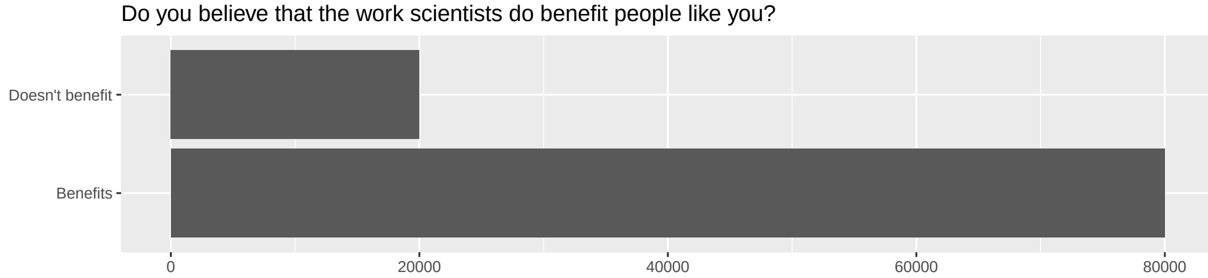
The Wellcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from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

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

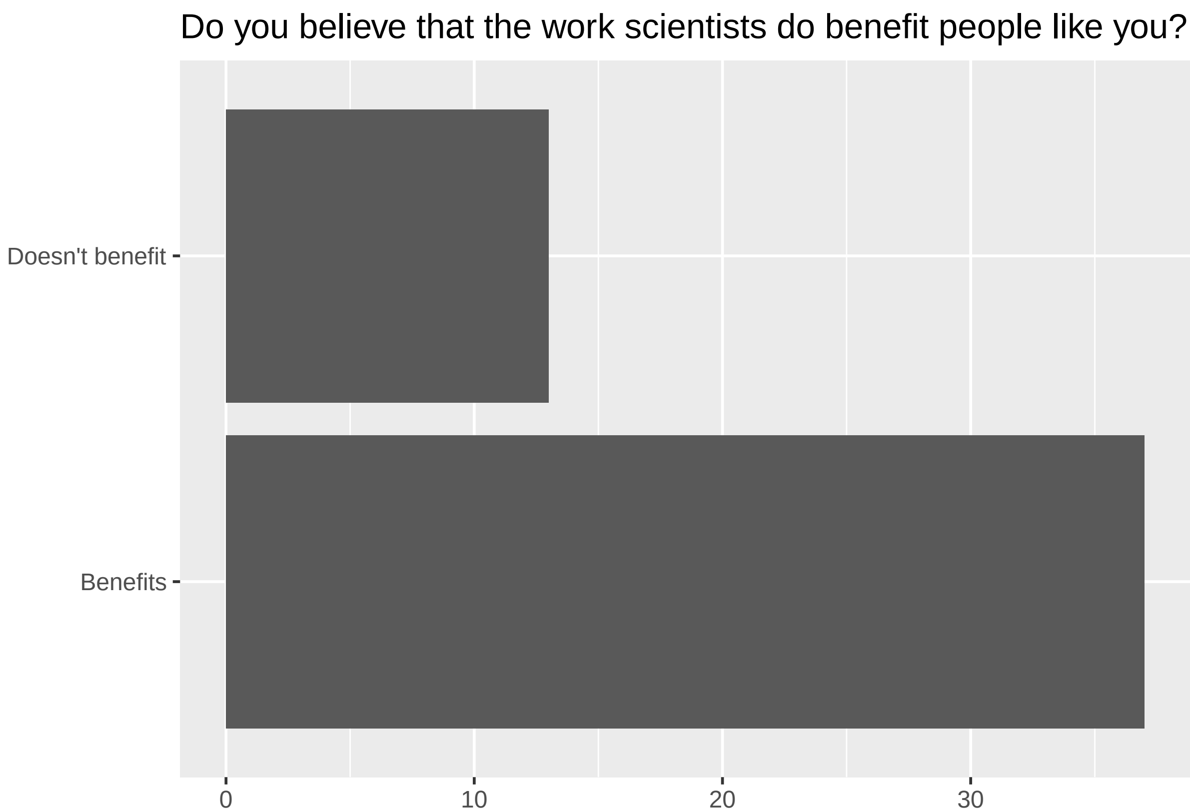
This command collects a simple random sample of size 50 from the `global_monitor` dataset, and assigns the result to `samp1`. This is similar to randomly drawing names from a hat that contains the names of all in the population. Working with these 50 names is considerably simpler than working with all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

Sample: samp1

```r
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()
```



Check the summary statistics of samp1

```r
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
```

The distribution of responses for the sample and population are similar. However, their proportions are different. For example, the proportion, p for "Benefits" is 0.8 for the population while the proportion, p_hat for "Benefits" is 0.74

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.

Solution 2:

I would not expect the sample proportion to match the sample proportion of another student's sample because the samples are different and the sampling is random. However, I would expect the proportion to be similar.

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:

Take second sample:

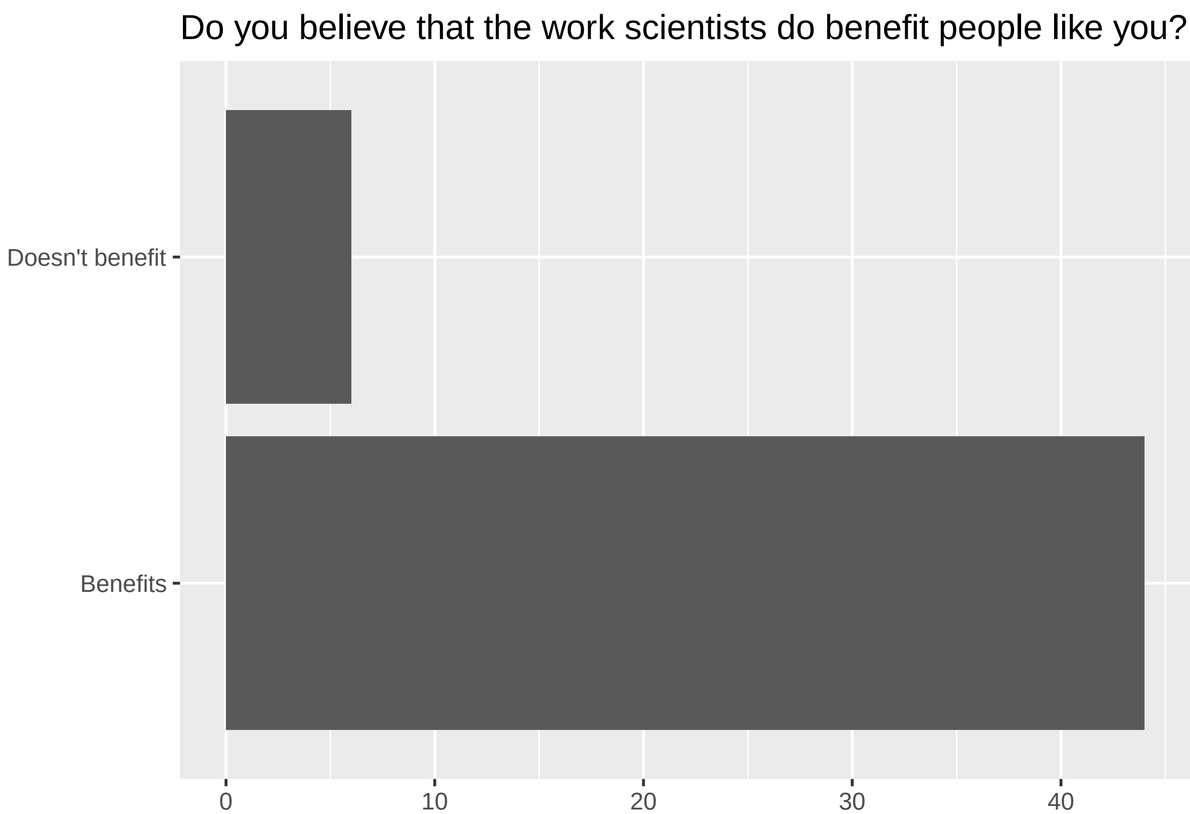
```
set.seed(110)
samp2 <- global_monitor %>%
  sample_n(50)

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           44   0.88
## 2 Doesn't benefit     6   0.12
```

Plot bar chart

```
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 proportion for samp2 is 0.88 while that for samp1 is 0.74 and they are different from each other. If we took two more samples of size 100 and 1000, the sample of size 1000 will provide a more accurate estimate of the population proportion since it is of higher magnitude. Larger sample sizes tend to provide results that are 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:

Generate the samples

```
set.seed(103)
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"
  )
```

Check the summary statistics

```
summary(sample_props50)
```

```
##    replicate     scientist_work           n              p_hat
##  Min.   :    1   Length:14999       Min.   : 1.00   Min.   :0.0200
##  1st Qu.: 3750   Class :character   1st Qu.: 8.00   1st Qu.:0.1600
##  Median : 7501   Mode  :character   Median :10.00   Median :0.2000
##  Mean   : 7501                      Mean   :10.03   Mean   :0.2006
##  3rd Qu.:11250                      3rd Qu.:12.00   3rd Qu.:0.2400
##  Max.   :15000                      Max.   :21.00   Max.   :0.4200
```

From the summary statistics of the sampling distribution, there are 15000 elements with a mean proportion, p_hat(Doesn't benefit) of 0.2006. The mean of p_hat(Benefit) = 1 - p_hat(Doesn't benefit) = 0.7994 which is very close to the mean of the population proportion. Also, from the histogram, the sampling distribution is uni-modal, symmetric and follows 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    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?

Solution 5:

```
set.seed(10)
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: 24 x 4
## # Groups:   replicate [24]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     5   0.5
##  2         2 Doesn't benefit     3   0.3
##  3         3 Doesn't benefit     3   0.3
##  4         4 Doesn't benefit     3   0.3
##  5         5 Doesn't benefit     5   0.5
##  6         6 Doesn't benefit     3   0.3
##  7         7 Doesn't benefit     4   0.4
##  8         8 Doesn't benefit     4   0.4
##  9         9 Doesn't benefit     2   0.2
## 10        10 Doesn't benefit     1   0.1
## # ... with 14 more rows
```

Check the summary statistics

```
summary(sample_props_small)
```

```
##    replicate     scientist_work           n               p_hat       
##  Min.   : 1.00   Length:24          Min.   :1.000   Min.   :0.1000  
##  1st Qu.: 6.75   Class :character   1st Qu.:2.000   1st Qu.:0.2000  
##  Median :12.50   Mode  :character   Median :3.000   Median :0.3000  
##  Mean   :12.54                      Mean   :2.792   Mean   :0.2792  
##  3rd Qu.:18.25                      3rd Qu.:3.250   3rd Qu.:0.3250  
##  Max.   :25.00                      Max.   :5.000   Max.   :0.5000  
```

There are 24 observations in this distribution with mean p_hat of 0.2792 and Each observation represents each sample.

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.)

Solution 6:

for size = 10 and 5,000 simulations: Mean = 0.22, SE = 0.11;
for size = 50 and 5,000 simulations: Mean = 0.2, SE = 0.06;
for size = 100 and 5,000 simulations: Mean = 0.2, SE = 0.04;

Each observation represents a sample. As the sample size increases, the mean of the sampling distribution converges to 0.2 while the standard error decreases significantly. Also, as sample size increases, the sampling distribution becomes more symmetric and appears to follow a more normal distribution.
If I increase the number of simulations, the mean of the sampling distribution does not change and the change in standard error is marginal. I think this is because 5,000 simulations in this case is already large enough for the sampling distribution values (Mean and SE) to converge to the true population values.

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?

Solutions 7:

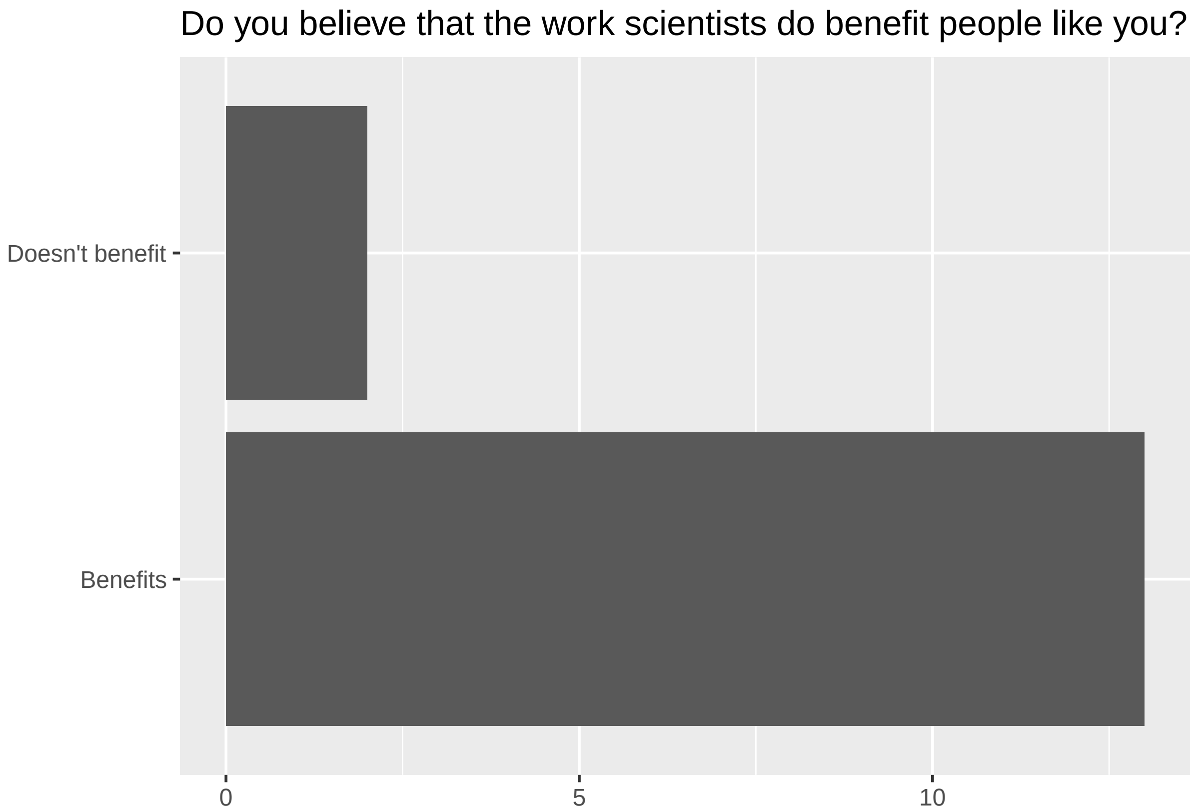
```
set.seed(114)
samp15 <- global_monitor %>%
  sample_n(15)

samp15 %>%
  count(scientist_work) %>%
  mutate(p_hat2 = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat2
##   <chr>           <int>  <dbl>
## 1 Benefits           13  0.867
## 2 Doesn't benefit     2  0.133
```

Plot bar chart

```
ggplot(samp15, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```



Using this sample, the best point estimate of the population proportion of people who think the work scientists do enhances their lives is 0.867

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:

Generate the sampling distribution

```
set.seed(111)
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")
```

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

```r
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

```r
summary(sample_props15)
```

```
##    replicate      scientist_work           n           p_hat
##  Min.   :   1.0   Length:2000        Min.   : 6   Min.   :0.4000
##  1st Qu.: 500.8   Class :character   1st Qu.:11   1st Qu.:0.7333
##  Median :1000.5   Mode  :character   Median :12   Median :0.8000
##  Mean   :1000.5                      Mean   :12   Mean   :0.7998
##  3rd Qu.:1500.2                      3rd Qu.:13   3rd Qu.:0.8667
##  Max.   :2000.0                      Max.   :15   Max.   :1.0000
```

Using this sample, I would guess the true population proportion to be 0.8. The calculations shows that the population proportion of people who think the work scientists do enhances their lives is 0.7998

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:

Generate the sampling distribution

```r
set.seed(111)
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")
```

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

```r
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

```
summary(sample_props150)
```

```
##    replicate      scientist_work           n             p_hat
##  Min.   :   1.0   Length:2000        Min.   :104   Min.   :0.6933
##  1st Qu.: 500.8   Class :character   1st Qu.:117   1st Qu.:0.7800
##  Median :1000.5   Mode  :character   Median :120   Median :0.8000
##  Mean   :1000.5                      Mean   :120   Mean   :0.7999
##  3rd Qu.:1500.2                      3rd Qu.:123   3rd Qu.:0.8200
##  Max.   :2000.0                      Max.   :135   Max.   :0.9000
```

Using this sample, I would guess the true population proportion to be 0.8. However, the calculations shows that the population proportion of people who think the work scientists do enhances their lives is 0.7999 which is very similar to the sampling distribution of size 15

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 with the smaller spread is the sample with a greater size. This basically means that as we increase the sample size, the spread of the data decreases and it gets closer to the actual population parameter. Hence, if I'm concerned with making estimates that are more often close to the true value, I would prefer a sampling distribution with a larger size and smaller spread.
