Foundations for statistical inference - Sampling distributions

In this assignment, 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.

Getting Started

Load packages

In this assignment, 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.

```r
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 assignment, 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.

```r
# This code is making a dataframe called global monitor.

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.

```r
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.

```
# This first line counts the science work benefiet and saves it as a new data frame. the letter n is us
g_monitor_count <- count(x = global_monitor, scientist_work, name ="count")

# This takes the data frame and adds a proportion variable
mutate(g_monitor_count, proportion = count /sum(count))
```

```
## # A tibble: 2 x 3
##   scientist_work  count proportion
##   <chr>           <int>      <dbl>
## 1 Benefits        80000        0.8
## 2 Doesn't benefit 20000        0.2
```

The unknown sampling distribution

In this assignment, 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. For the remainder of this R Assignment pretend we do not have access to the entire population and are only able to create a sample.

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 <-  sample_n(global_monitor, 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.

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 proportion.

```
# This first line counts the science work benefiet and saves it as a new data frame. the letter n is us
g_monitor_count_samp <- count(x = samp1, scientist_work, name ="count")

# This takes the data frame and adds a proportion variable
mutate(g_monitor_count_samp, proportion = count /sum(count))
```

1. Describe the distribution of responses in this sample. That is how has the proportion changed? Why does this seem reasonable?

Depending on which 50 people you selected, your sample proportion (p_hat) could be a bit above or a bit below the true population proportion of 0.20. 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.

1. Would you expect the sample proportion to match the sample proportion of another student's sample? Why, or why not? Ask a student team to confirm your answer.
2. Write the code to 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?

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 proportion this way. The distribution of sample proportions, called the *sampling distribution (of the proportion)*, can help you understand this variability. In this assignment, 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`. This collection of sample statistics is called the *sampling distribution.*

Note that we specify that `replace = TRUE` since sampling distributions are constructed by sampling with replacement.

```
# Here was use a special operator called the pipe |>, which takes the output of the previous line and p

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.

1. 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.

Interlude: Sampling distributions

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

With the `rep_sample_n` function, we were able to simulate 15000 samples with 15000 statistics and visualize them in the histogram above. (Yes that is 15000 samples in that graph)

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

This is similar to what happens in real life, except usually pollsters only collect one sample and statisticians do some math to guess at the population proportion, because they know it will have the normal shape above.

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.

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 proportion.

Graded Questions:

The following questions will be peer graded.

The code below was copied from above and will produce a histogram based on the sample size and number of samples simulated. You task is to explore what happens to the center, spread, and shape of the distribution as the sample size increases. Change the code and write your answer at the bottom. Then knit and upload this document to R_assignment.

```r
# Change this number and run the chunk.
sample_size = 10
repetitions = 1000

sample_props <- global_monitor %>%
                    rep_sample_n(size = sample_size, reps = repetitions, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")

ggplot(data = sample_props, 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"
  )
```

Questions:

It is my hope that you have a sense of the central limit theorem at this point. Some of the new code can see overwhelming, but don't worry too much about it.

There are only 3 questions for this R Assignment.

1. As the sample size increases the center of the sampling distribution gets closer to the true proportion
2. As the sample size increases the variability of the sampling distribution decreases
3. As the sample size increases the shape becomes more normal