Foundations for statistical inference - Sampling distributions

William Jasmine

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)
library(gridExtra)
```

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?

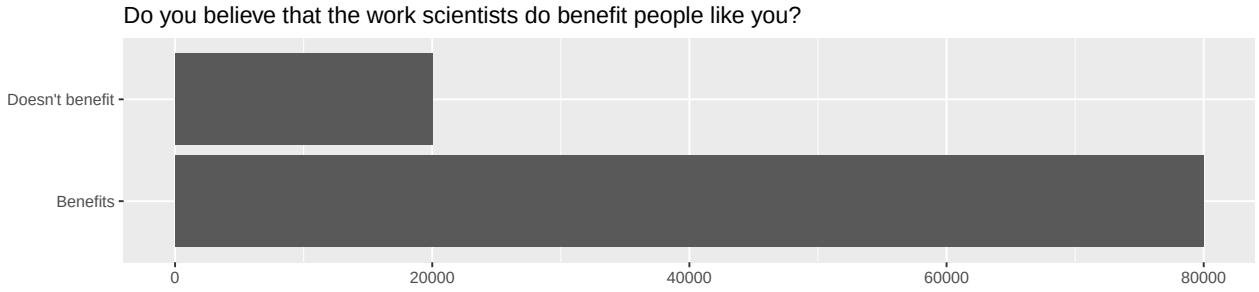
The Wellcome Global Monitor finds that 20% of people globally do not believe that the work scientists do benefits people like them. In this lab, you will assume this 20% is a true population proportion and learn about how sample proportions can vary from sample to sample by taking smaller samples from 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(seed)
samp1 <- global_monitor %>%
  sample_n(50)
```

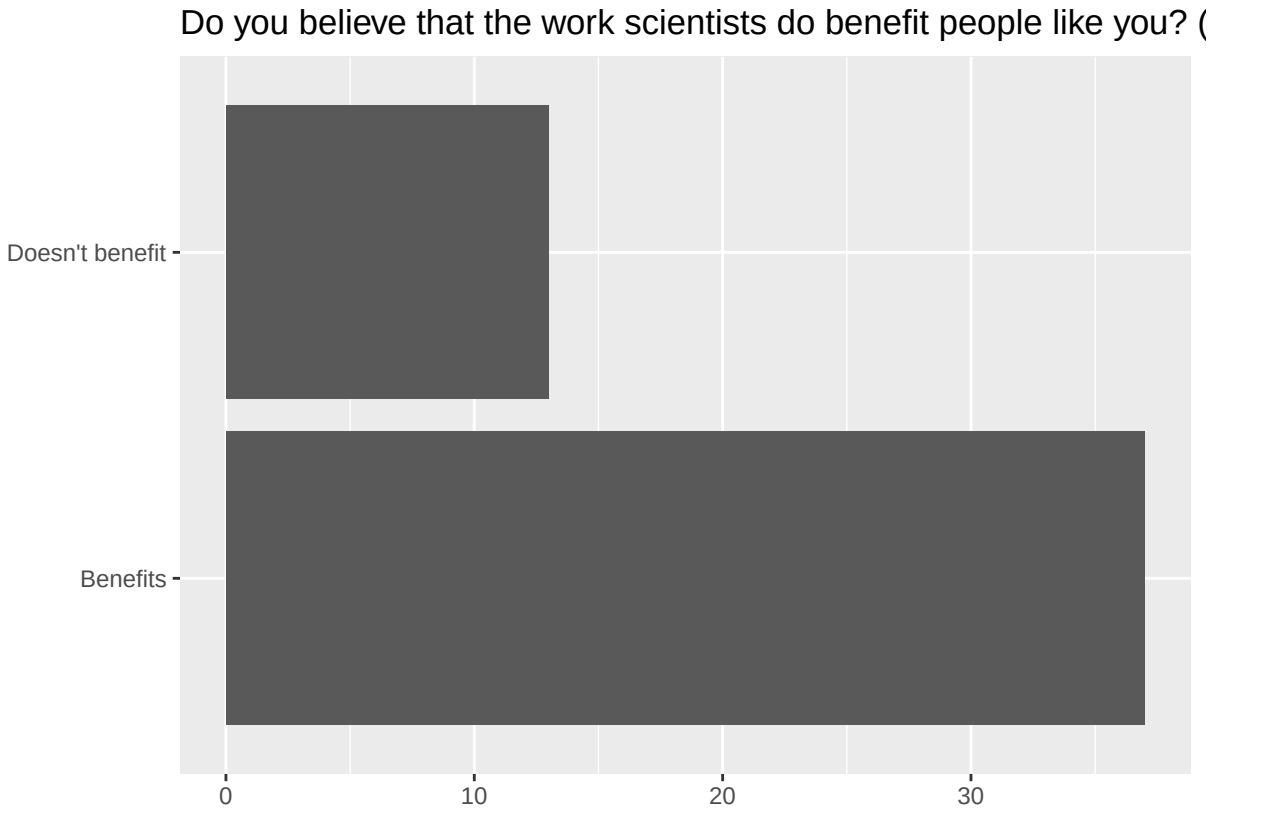
This command collects a simple random sample of size 50 from the `global_monitor` dataset, and assigns the result to `samp1`. This is similar to randomly drawing names from a hat that contains the names of all in the population. Working with these 50 names is considerably simpler than working with all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

WJ Response:

The below code block makes a histogram of the previously created `samp1` sample of `global_monitor`:

```
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you? (Sample)"
  ) +
  coord_flip()
```



Given that `samp1` has a size of 50, we would expect from the `global_monitor` probabilities for there to be 10 people who don't see the benefits of scientists, and 40 who do. Because of the random sampling, each category is close to these values but not exact. The exact metrics of `samp1` are shown below:

```
samp1 %>%
  group_by(scientist_work) %>%
    summarise(n_sample = n(), p_sample = n() / nrow(samp1))
```

```
## # A tibble: 2 x 3
##   scientist_work  n_sample p_sample
##   <chr>              <int>    <dbl>
## 1 Benefits              37     0.74
## 2 Doesn't benefit       13     0.26
```

As the above outputs show, the results are split 37/13 instead of 40/10.

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 sample 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.

WJ Response:

If multiple samples were taken, it might be the case that some of these samples would match exactly to the results of `samp1`, but they will likely all at least be similar. As the proportion of values strays away from 40/10, it will be less and less likely for multiple samples to obtain those values.

---

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?

WJ Response:

```
set.seed(seed+1)
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           41  0.82
## 2 Doesn't benefit     9  0.18
```

The code chunk above creates `samp2`, a different random sample from `global_monitor` using a different seed. As mentioned in problem 2, the proportions of this sample are close but not exactly equal to those present in `samp1` (41:9 as opposed to 37:13). The average of the two is in fact closer to the expected 40:10 split.

If the size of these samples were increased to 1,000 as opposed to 50 or 100, they would be better approximations of the true mean thanks to the Law of Large Numbers.

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.

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

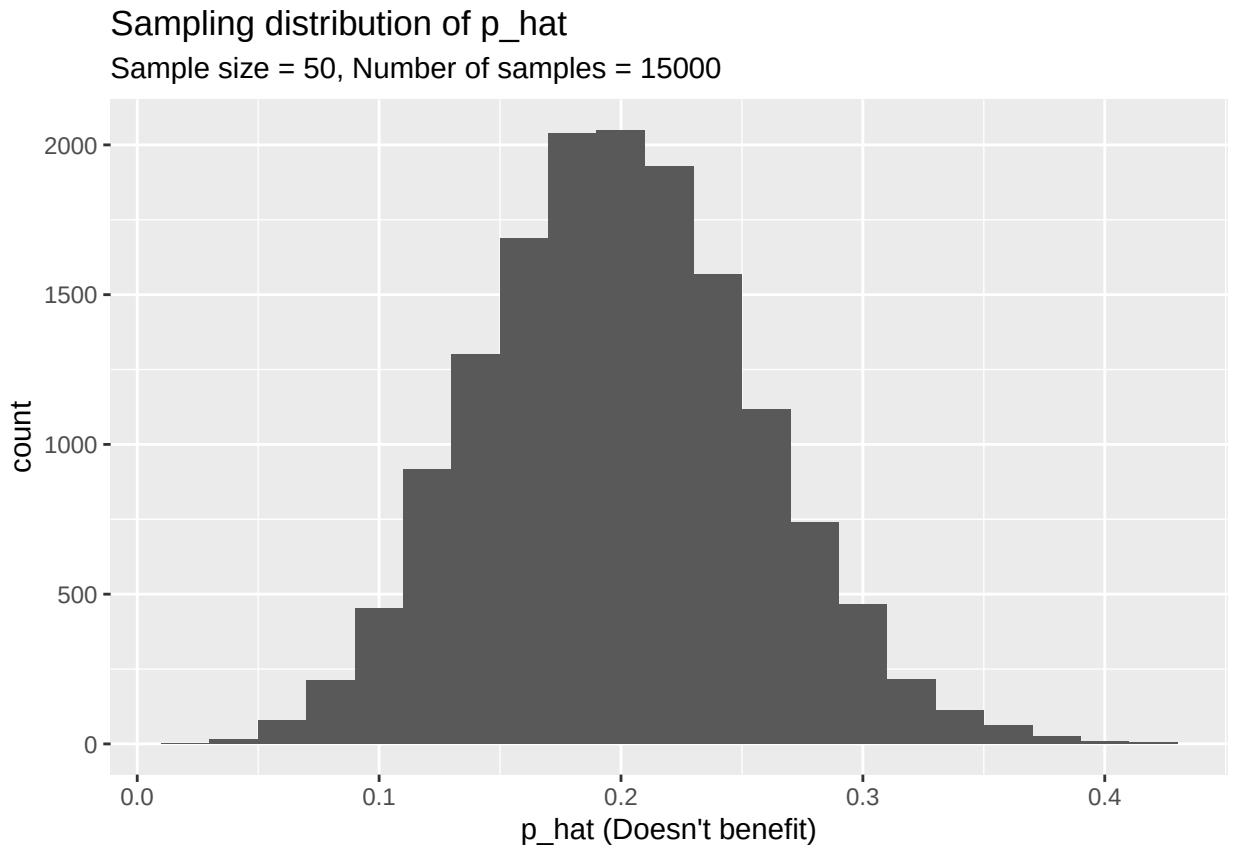
Next, you will review how this set of code works.

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.

WJ Response:

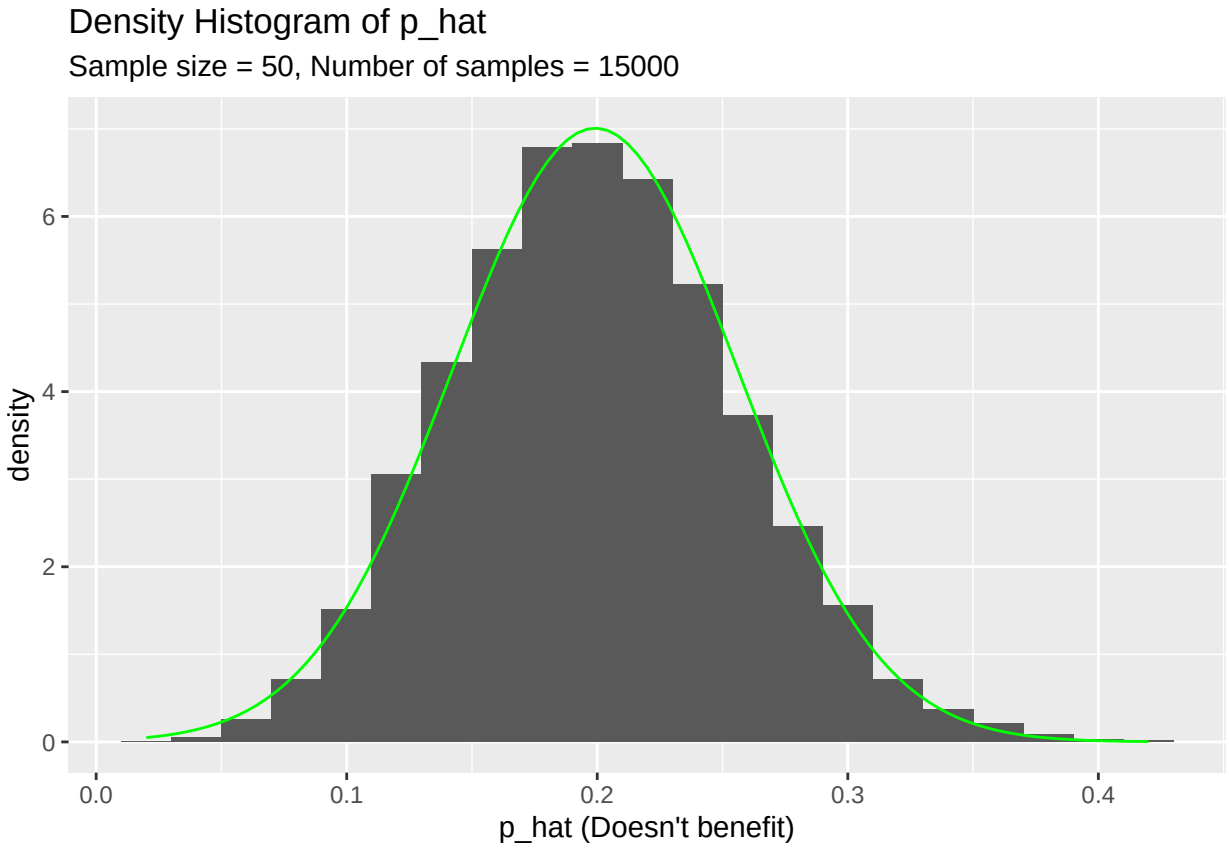
The size of the `sample_props50` dataframe is 15,000 due to the fact that the above code created 15,000 samplings of the `global_monitor` dataframe, and in every one of those samples there was at least one person (out of 50) who thought that scientists don't benefit them. The distribution of the proportions in `sample_props50` is plotted in a histogram below:

```
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"
  )
```



We can clearly see that the mean is centered around 0.2 (exactly the probability we'd expect for the percentage of people who do not feel scientists benefit them), and that the distribution looks normal. We can investigate this normality by plotting the density histogram of `sample_props50` above along with a normal distribution curve using `sample_props50`'s mean and standard deviation values:

```
ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_blank() +
  geom_histogram(aes(y = ..density..), binwidth = 0.02) +
  stat_function(fun = dnorm,
                args = c(mean = mean(sample_props50$p_hat),
                         sd = sd(sample_props50$p_hat)),
                col = "green") +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Density Histogram of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```



As is clear in the plot above, the density histogram almost perfectly approximates the 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    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?

WJ Response:

The code chunk below creates the `sample_props_small` dataframe, which takes 25 random samples of size 10 from the `global_monitor` dataframe.

```
set.seed(seed=seed)

sample_props_tmp <- global_monitor %>%
  rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
    count(scientist_work) %>%
      mutate(p_hat = n /sum(n))

sample_props_small <- sample_props_tmp %>%
  filter(scientist_work == 'Doesn\'t benefit')

print(sample_props_tmp)
```

```
## # A tibble: 47 x 4
## # Groups:   replicate [25]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Benefits            7   0.7
##  2         1 Doesn't benefit     3   0.3
##  3         2 Benefits            7   0.7
##  4         2 Doesn't benefit     3   0.3
##  5         3 Benefits            8   0.8
##  6         3 Doesn't benefit     2   0.2
##  7         4 Benefits           10   1
##  8         5 Benefits            5   0.5
##  9         5 Doesn't benefit     5   0.5
## 10         6 Benefits            9   0.9
## # ... with 37 more rows
```

We can see in the output above that there are only 22 rows despite there being 25 samples taken in the the code above. This is due to the fact that there are three samples in which not a single person believed that scientists do not benefit them on a day to day basis. Given that the sample size was reduced to 10 (as opposed to 50 for the `sample_props50` dataframe), this kind of outcome is much more likely. This is confirmed by looking at the `sample_props_tmp` dataframe, which was created as an intermediate step before the filtering that produced `sample_props_small`:

```
sample_props_tmp %>%
  filter(p_hat == 1)
```

```
## # A tibble: 3 x 4
## # Groups:   replicate [3]
##   replicate scientist_work     n p_hat
##       <int> <chr>          <int> <dbl>
## 1         4 Benefits          10     1
```

```
## 2         20 Benefits           10      1
## 3         21 Benefits           10      1
```

The code above shows that there are 3 rows that have where `scientist_work` = "Benefits" and `p_hat` = 1, meaning that there are 3 samples that contained no people who thought scientists did not benefit them. As such, the rows in `sample_props_small` represent the percentage of people in each sample that thought scientist's work does not benefit them, for only the samples in which there was at least 1 person that thought so.

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.)

WJ Response:

Each of the observations in the sampling distributions produced by the Shiny app represent the proportion of people in a single simulation who do not think that scientists benefit their lives on a day to day basis. Each simulation in this case consists of n people (controlled by the "Sample Size" input in the app) and is repeated N times (controlled by the "Number of Samples" input in the app). These values are then used to create the histogram.

It is clear from using different values of n, that increasing the sample size of each sample has the following effects:

- Centers the histogram closer to the expected mean (in this case, 0.2).
- Reduces the standard error and spread of the histogram (the bars become more tightly centered around the mean).
- Causes the distribution to look increasingly more normal.

As for changing the value of N: the histograms do not shed that much information for low values of N, but once N is greater than around 100 samples, it does not seem to change the shape of the distribution very much.

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?

WJ Response:

The below code creates a sample of size 15 from `global_monitor` and uses it to estimate the percentage of people who do believe science benefits them on a day-to-day basis:

```
set.seed(seed)
samp3 <- global_monitor %>%
  sample_n(15)

samp3 %>%
  filter(scientist_work=='Benefits') %>%
    nrow() / nrow(samp3)
```

```
## [1] 0.7333333
```

Based on the above, we would estimate that ~73.3% percent of the population believes that scientists' work does benefit their day to day lives.

---

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.

WJ Response:

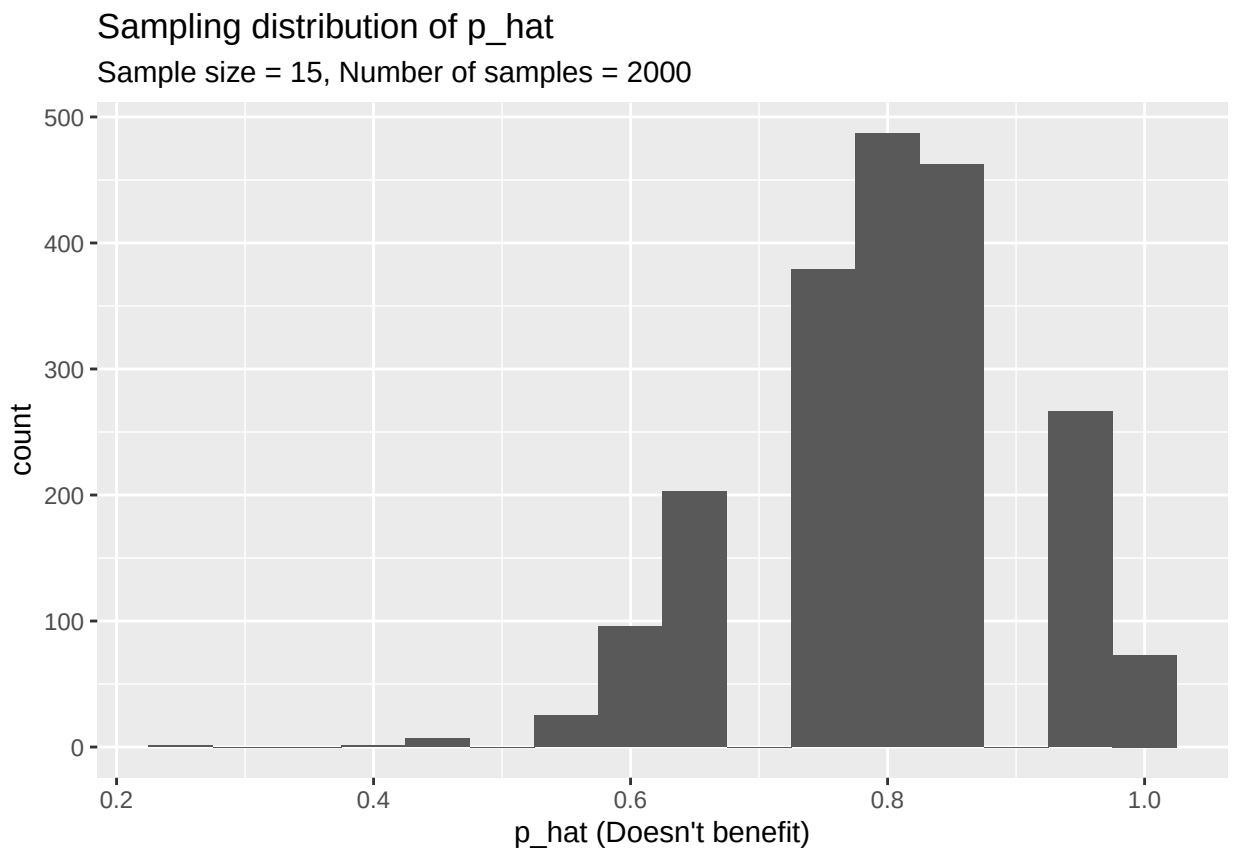
The code chunk below creates `sample_props15` a dataframe containing the results of 2,000 simulations that records the percentage of 15 randomly chosen people who think scientists benefit their day to day lives:

```
set.seed(seed)
sample_props_tmp2 <- global_monitor %>%
                     rep_sample_n(size = 15, reps = 2000, replace = TRUE)

sample_props15 <- sample_props_tmp2 %>%
                     count(scientist_work) %>%
                     mutate(p_hat = n /sum(n)) %>%
                     filter(scientist_work == "Benefits")
```

The sampling distribution contained within the `sample_props15` dataframe is shown in a histogram below:

```
ggplot(data = sample_props15, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.05) +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 15, Number of samples = 2000"
  )
```



Given that the distribution shown above is centered at 0.8 and the data looks normal, we can estimate that 80% of people in our population believe that the work of scientists benefit their day to day lives. The below chunk takes a closer look into this, by reporting the population proportions of the entire set of samples from `sample_props15`:

```
sample_props_tmp2 %>%
  group_by(scientist_work) %>%
    summarise(n = n(), p = n()/nrow(sample_props_tmp2))
```

```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits        23991 0.800
## 2 Doesn't benefit  6009 0.200
```

We can see that based of of the 30,000 (15 x 2,000) people that were randomly sampled (with replacement), that almost exactly 80 percent of them thought that scientists' work does benefit them.

---

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?

WJ Response:

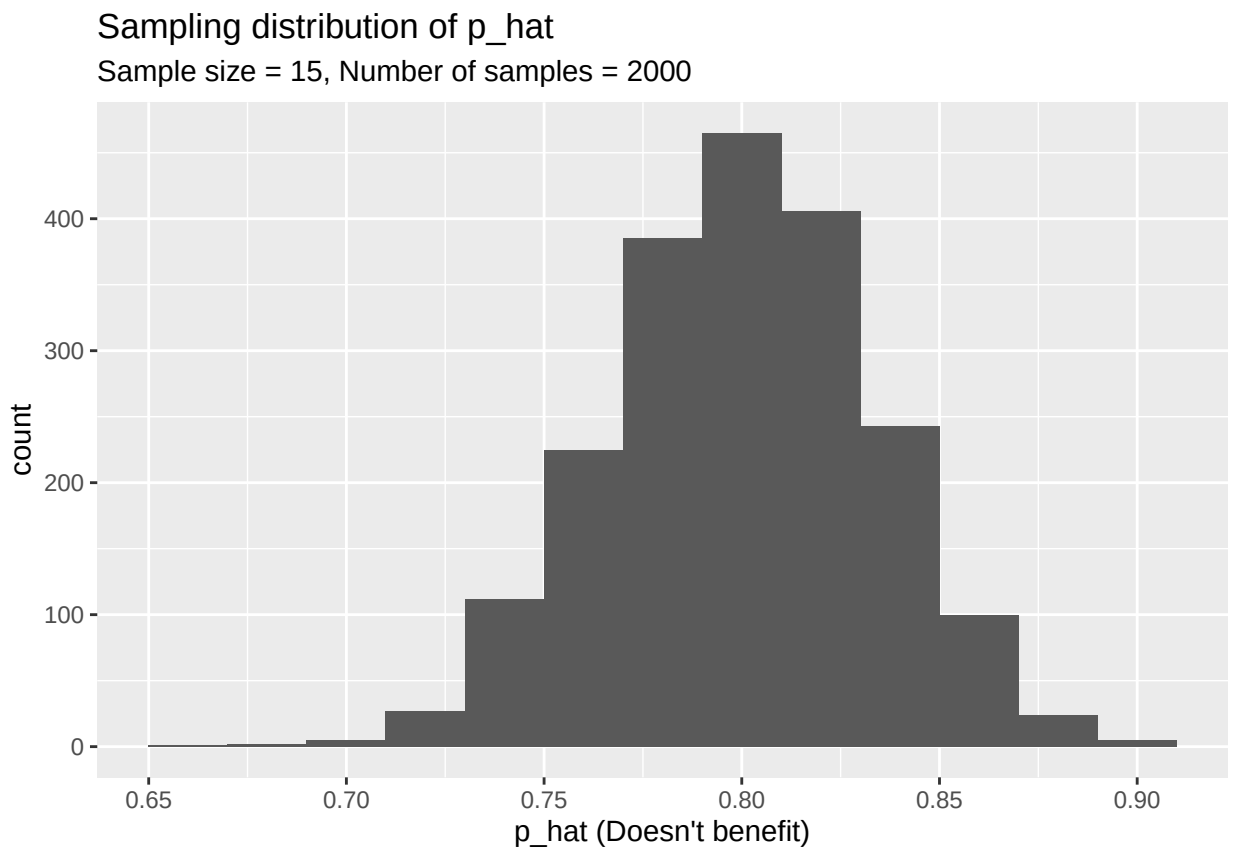
The code chunk below creates `sample_props150` a dataframe containing the results of 2,000 simulations that records the percentage of 150 randomly chosen people who think scientists benefit their day to day lives:

```
set.seed(seed)
sample_props_tmp2 <- global_monitor %>%
                     rep_sample_n(size = 150, reps = 2000, replace = TRUE)

sample_props150 <- sample_props_tmp2 %>%
                     count(scientist_work) %>%
                     mutate(p_hat = n /sum(n)) %>%
                     filter(scientist_work == "Benefits")
```

The sampling distribution contained within the `sample_props15` dataframe is shown in a histogram below:

```
ggplot(data = sample_props150, 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 = 15, Number of samples = 2000"
  )
```



Given that the distribution shown above is centered as 0.8 and the data looks normal, we can estimate that 80% of people in our population believe that the work of scientists benefit their day to day lives. This distribution is even more tightly bound around 0.8 compared to the histogram created previously with sample sizes of 15, and looks almost perfectly symmetric/normal. The below chunk takes a closer look at the population proportions of the entire set of samples from `sample_props150`:

```
sample_props_tmp2 %>%
  group_by(scientist_work) %>%
    summarise(n = n(), p = n()/nrow(sample_props_tmp2))
```

```
## # A tibble: 2 x 3
##   scientist_work       n     p
##   <chr>            <int> <dbl>
## 1 Benefits        239926 0.800
## 2 Doesn't benefit  60074 0.200
```

We can see that based of of the 300,000 (150 x 2,000) people that were randomly sampled (with replacement), that almost exactly 80 percent of them thought that scientists' work does benefit them. In addition, the proportions are even closer to the expected 80/20 split compared to when the sample size was only 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?

WJ Response:

The sampling distribution created from the samples that had size 150 most definitely has less spread and more closely approximates the normal distribution thanks to the central limit theorem. As such, in order to make estimates that are closer to the true value, it is always beneficial to have a larger sample size.
