test

Jack Wright

10/4/2020

```
library(tidyverse)
library(openintro)
library(infer)
```

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 summarising 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.

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.

1. 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.

I would expect all of our sample proportions to be different, since it is a random sample. Of course the distribution of our samples are most likely to be around the mean of the sampling distribution. I confirmed this with my group.

1. 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?

```
samp2 <- global_monitor %>%
  sample_n(50)

samp2_p_hat <- samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Doesn't benefit") %>%
  pull(p_hat) %>%
  round(2)
```

I would expect the larger sample to be more accurate.

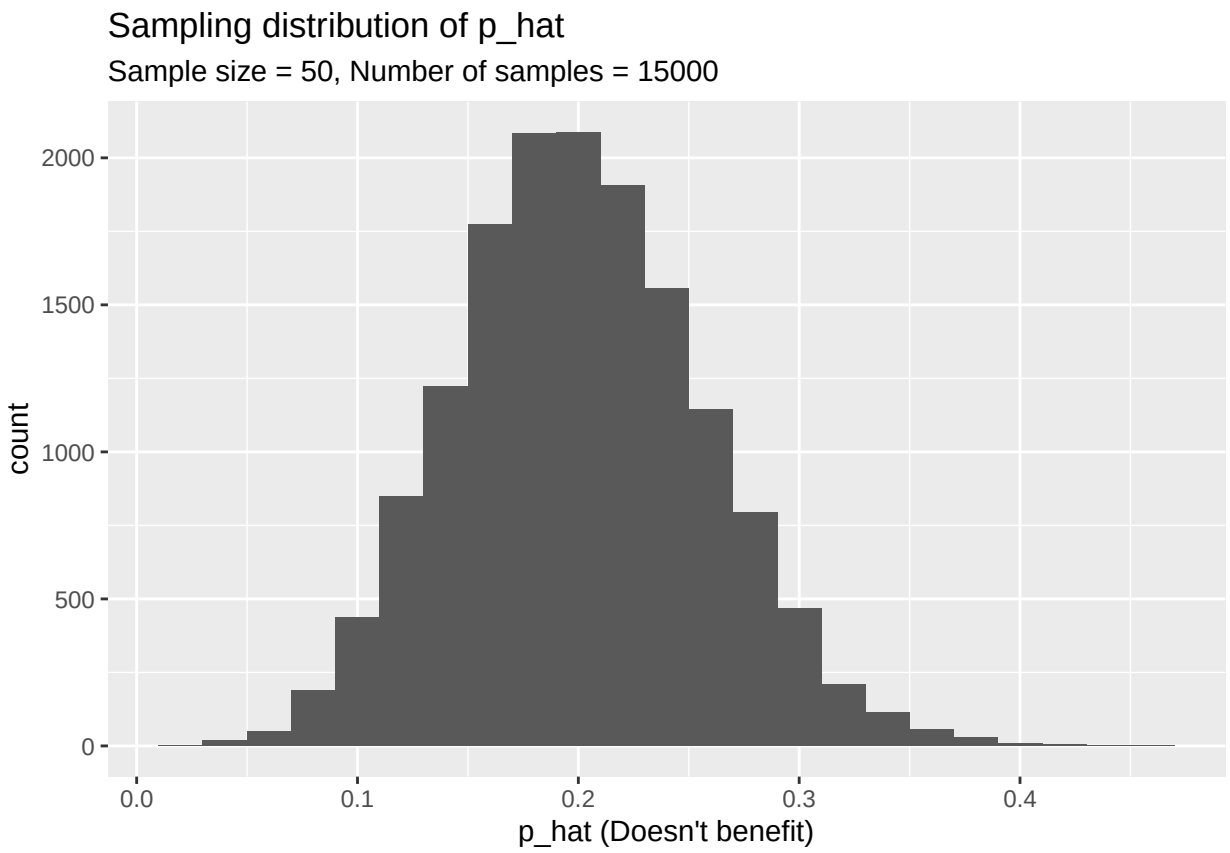
1. How many elements are there in `sample_props50`? Describe the sampling distribution, and be sure to specifically note its center. Make sure to include a plot of the distribution in your answer.

There are 15,000 observations in `sample_props50`. The sampling distribution looks normal.

```
samp50_p_hat<- mean(sample_props50$p_hat)
SE<-sd(sample_props50$p_hat)/1500^.2
print(cat("sampling mean=",round(samp50_p_hat,3), "and the standard error=",round(SE,3)))
```

```
## sampling mean= 0.2 and the standard error= 0.013NULL
```

```
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"
  )
```



1. 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?

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

There are 23 observations, (an observation of n=0 does not show up, so 25 samples) in this data frame, and each object represents a sample of 10 values from global_monitor.

1. 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.)

Each observation in the sampling distribution represents the mean of a sample. As the number of samples increases, the mean gets closer to the mean of the dataset and the standard error decreases. This makes sense because this sampling process is used to find the mean, and as you take more means, the accuracy increases (SE drops).

1. Take a sample of size 15 from the population and calculate the proportion of people in this sample who think the work scientists do enchances 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?

```
samp3 <- global_monitor %>%
  sample_n(15)

samp3_p_hat <- samp3 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Benefits") %>%
  pull(p_hat) %>%
  round(2)


cat("point estimate of the population proportion of people who think the work scientists do enchances th
```

```
## point estimate of the population proportion of people who think the work scientists do enchances thei
```

1. 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.

```
sample_props15<-global_monitor %>%
  rep_sample_n(size = 15, replace = TRUE, reps=2000) %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Benefits")
```

```r
p_hat<-round(mean(sample_props15$p_hat),2)

print(cat("point estimate of the population proportion of people who think the work scientists do encha
```

```
## point estimate of the population proportion of people who think the work scientists do enchances the
```

1. 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?

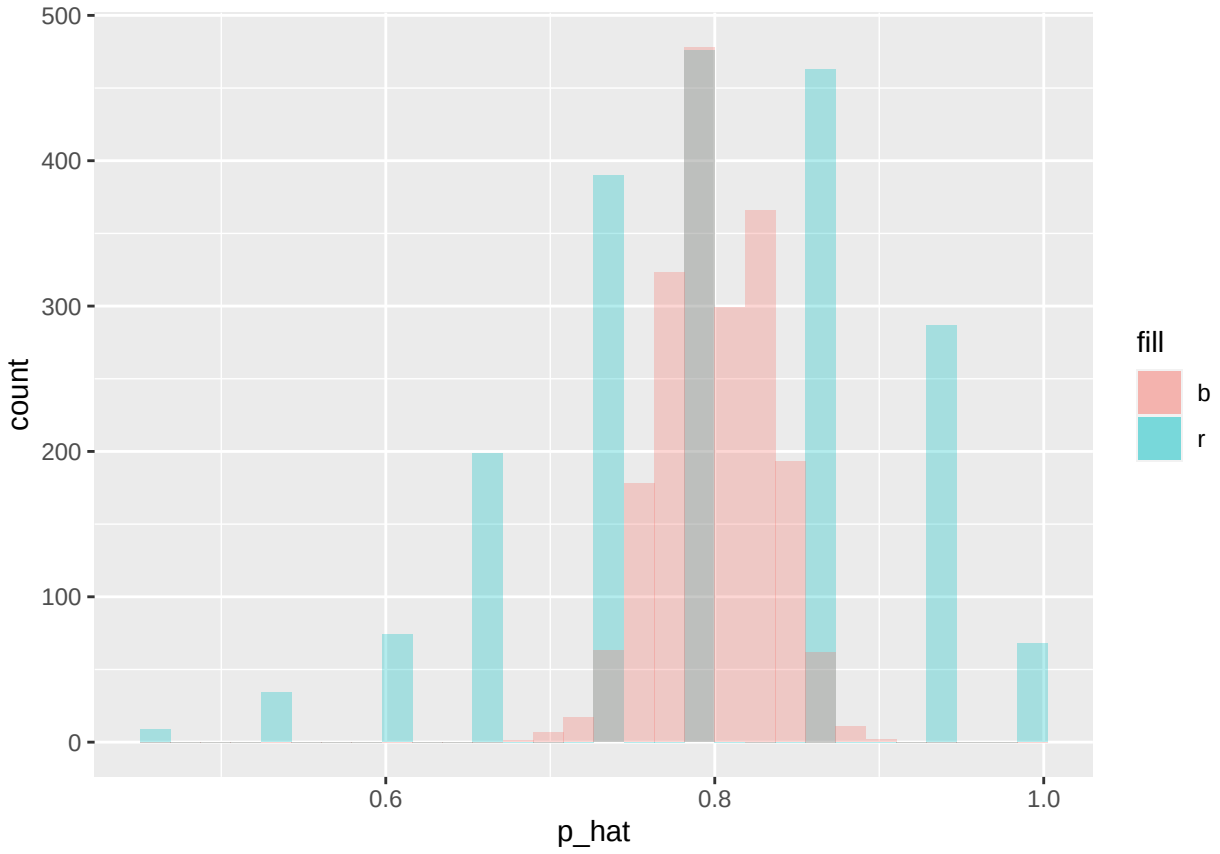
The shapes are both relatively normal, but the spread for the larger sample size is smaller. The means are the same.

```r
sample_props150<-global_monitor %>%
  rep_sample_n(size = 150, replace = TRUE, reps=2000) %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) %>%
  filter(scientist_work == "Benefits")

#ggplot(data = sample_props150, aes(x = p_hat)) +
  #geom_histogram(binwidth = 0.02)

ggplot()+
  geom_histogram(data=sample_props15,aes(x=p_hat, fill="r"),alpha=.3)+
  geom_histogram(data=sample_props150,aes(x=p_hat, fill="b"),alpha=.3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1. 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?

The more accurate you need to be, the smaller the spread should be.

```
## point estimate of the population proportion of people who think the work scientists do enchances the
```