Lab5a - Sampling Distributions

Keith Colella

2023-03-12

Setup

Config

```
library(tidyverse)
library(openintro)
library(infer)
set.seed(789)
```

Data

```
global_monitor <- tibble(
  scientist_work = c(rep("Benefits", 80000), rep("Doesn't benefit", 20000))
)
```

Exercise 1

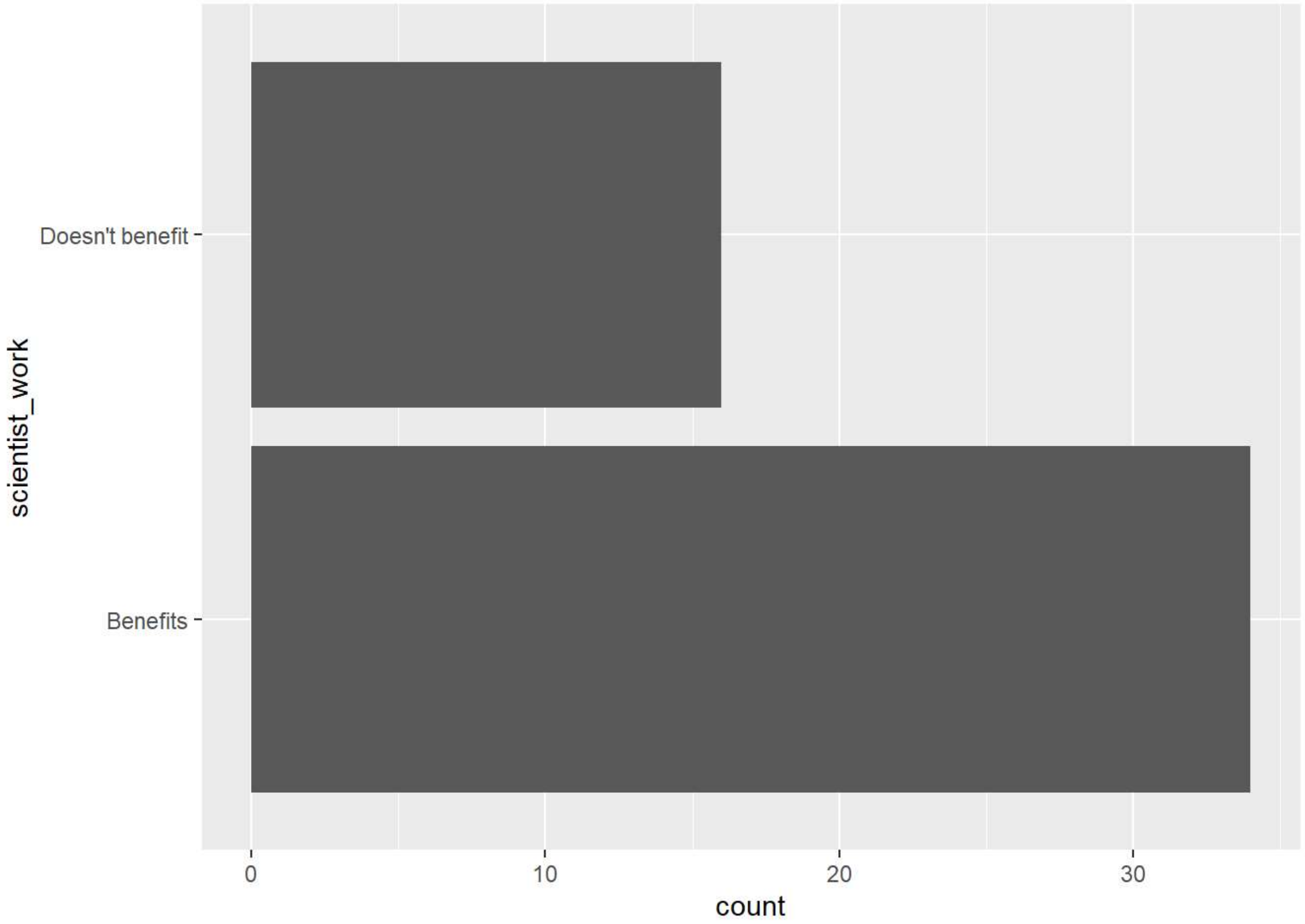
Question

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.

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

Response

```
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  coord_flip()
```



```
as.data.frame(table(samp1)) %>%
  mutate(Proportion = Freq / sum(Freq))
```

```
##    scientist_work Freq Proportion
## 1        Benefits   34       0.68
## 2 Doesn't benefit   16       0.32
```

The distribution of responses in our sample differs significantly from the population distribution. In our sample, roughly a third (~32%) of respondents do not believe scientific research benefits them, representing 1.5x the proportion seen in the broader population.

Exercise 2

Question

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.

Response

I would not expect them to match (unless our random seeds matched, but we'll discount that possibility for now). Random sampling should not result in the same observations being sampled across separate experiments. So we should expect to sample different subsets of individuals, driving different sample proportions. We should expect some degree of consistency, but not too much. A sample size of 50 to represent a population of 100,000 is likely to introduce significant variance.

Exercise 3

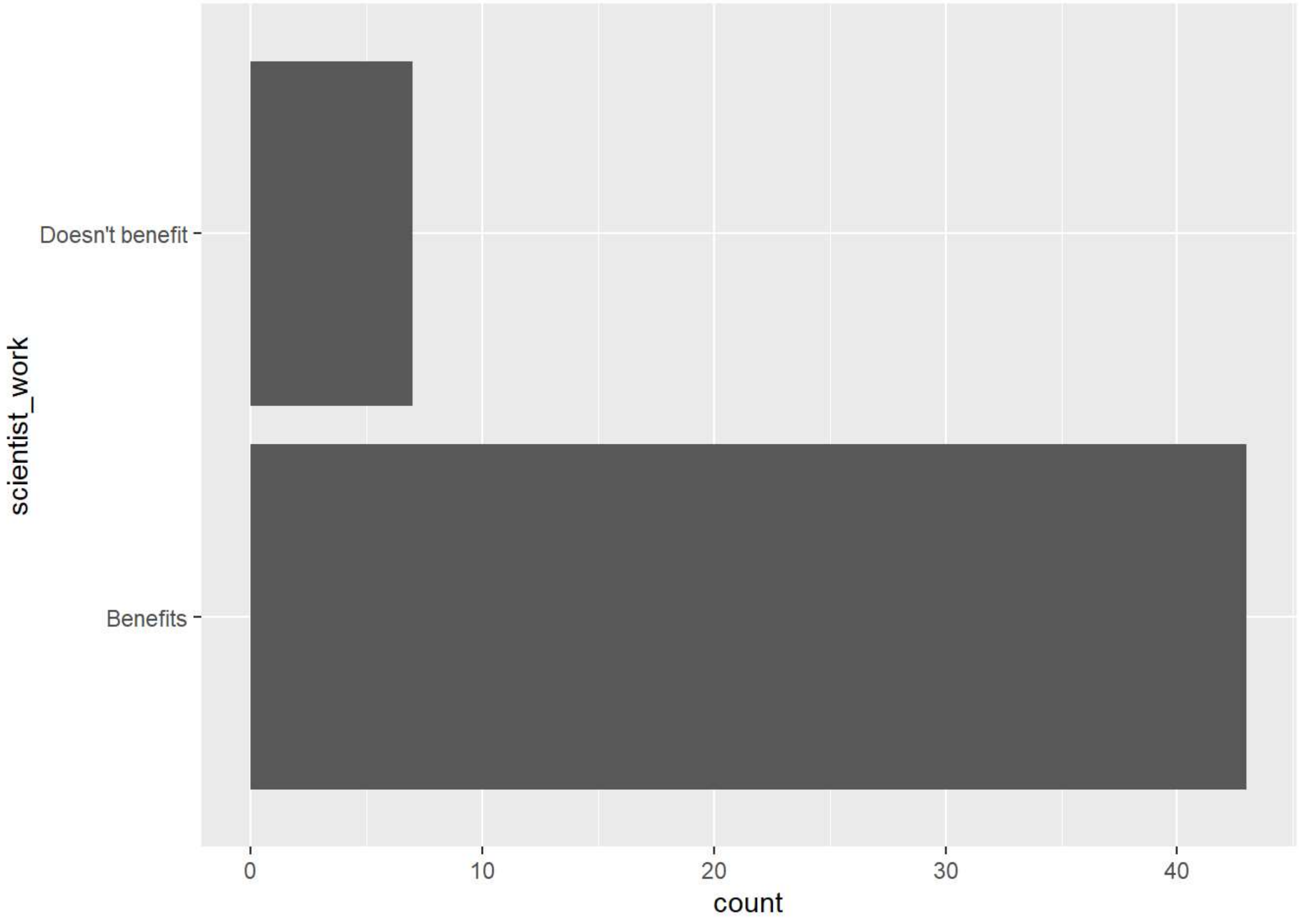
Question

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?

Response

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

ggplot(samp2, aes(x = scientist_work)) +
  geom_bar() +
  coord_flip()
```



```
as.data.frame(table(samp2)) %>%
  mutate(Proportion = Freq / sum(Freq))
```

```
##    scientist_work Freq Proportion
## 1        Benefits   43       0.86
## 2 Doesn't benefit    7       0.14
```

In our second sample, we actually see $\hat{p}$ much closer to p at 7 out of 50 or 14%. If we were to take two additional samples of 100 and 1000, I would expect their sample means to become increasingly closer to the population means. Let’s test it out!

```
sample_sizes <- c(100,1000)

for (n in sample_sizes) {
  samp <- global_monitor %>% sample_n(n)
  as.data.frame(table(samp)) %>%
    mutate(Proportion = Freq / sum(Freq)) %>%
    print()
}
```

```
##    scientist_work Freq Proportion
## 1        Benefits   74       0.74
## 2 Doesn't benefit   26       0.26
##    scientist_work Freq Proportion
## 1        Benefits  809      0.809
## 2 Doesn't benefit  191      0.191
```

As expected, we see the sample proportions converge towards 20% as the sample size increases.

Exercise 4

Question

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.

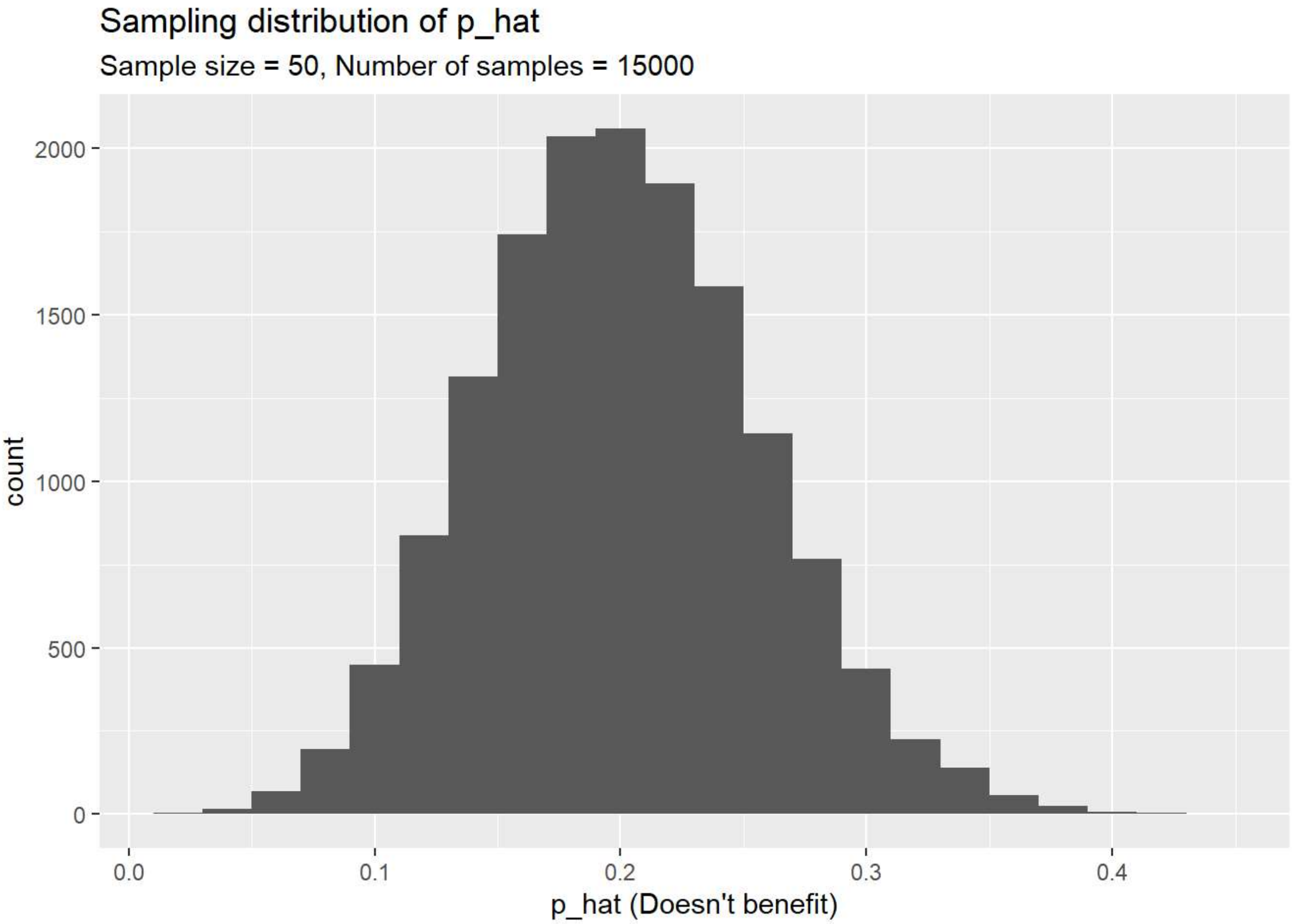
```
sample_props50 <- global_monitor %>%
                    rep_sample_n(size = 50, reps = 15000, replace = TRUE) %>%
                    count(scientist_work) %>%
                    mutate(p_hat = n /sum(n)) %>%
                    filter(scientist_work == "Doesn't benefit")
```

Response

```
head(sample_props50)
```

```
## # A tibble: 6 × 4
## # Groups:   replicate [6]
##   replicate scientist_work      n p_hat
##       <int> <chr>           <int> <dbl>
## 1         1 Doesn't benefit     9  0.18
## 2         2 Doesn't benefit    14  0.28
## 3         3 Doesn't benefit    12  0.24
## 4         4 Doesn't benefit     9  0.18
## 5         5 Doesn't benefit    10  0.2
## 6         6 Doesn't benefit    11  0.22
```

```
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"
  )
```



The `sample_props50` dataframe contains 15,000 observations (each comprised of 50 samples) of four variables, including the trial number, the response on which we're focused ("Doesn't benefit"), the number of respondents who gave that response in each trial, and the resulting sample proportion or $\hat{p}$, calculated as the number of respondents who responded "Doesn't benefit" divided by 50.

Using the same plot as that found in the lab, the resulting distribution is very nearly normal. The distribution's mean is very clearly centered around 0.20, which matches the population proportion. This provides very clear support for the Central Limit Theorem.

Exercise 5

Question

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?

Response

```
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 == 'Benefits') %>%
  mutate(doesnt_benefit = 10 - n,
         p_hat = 1 - p_hat,
         sample_size = 10) %>%
  select(replicate, sample_size, doesnt_benefit, p_hat)

print(sample_props_small)
```

```
## # A tibble: 25 × 4
## # Groups:   replicate [25]
##    replicate sample_size doesnt_benefit p_hat
##        <int>       <dbl>          <dbl> <dbl>
##  1         1          10              3   0.3
##  2         2          10              1   0.1
##  3         3          10              0   0
##  4         4          10              0   0
##  5         5          10              2   0.2
##  6         6          10              3   0.3
##  7         7          10              0   0
##  8         8          10              3   0.3
##  9         9          10              0   0
## 10        10          10              2   0.2
## # … with 15 more rows
```

The dataframe contains 25 observations, each capturing the number of "successes" (i.e. number of respondents who do not find benefits) for each trial, as well as the resulting sample proportion.

One thing that became clear while constructing the data frame was that the `rep_sample_n` function does *not* return rows when n = 0. In other words, if no "Doesn't benefit" responses are sampled in a given iteration, then only the results for "Benefits" are added for the dataframe. As a result, I had to initially focus on "Benefits" results, then calculate the complement to back out “Doesn’t benefit” results. Otherwise, our dataframe would have had less than 25 observations, which might skew any analysis of aggregated results.

Exercise 6

Question

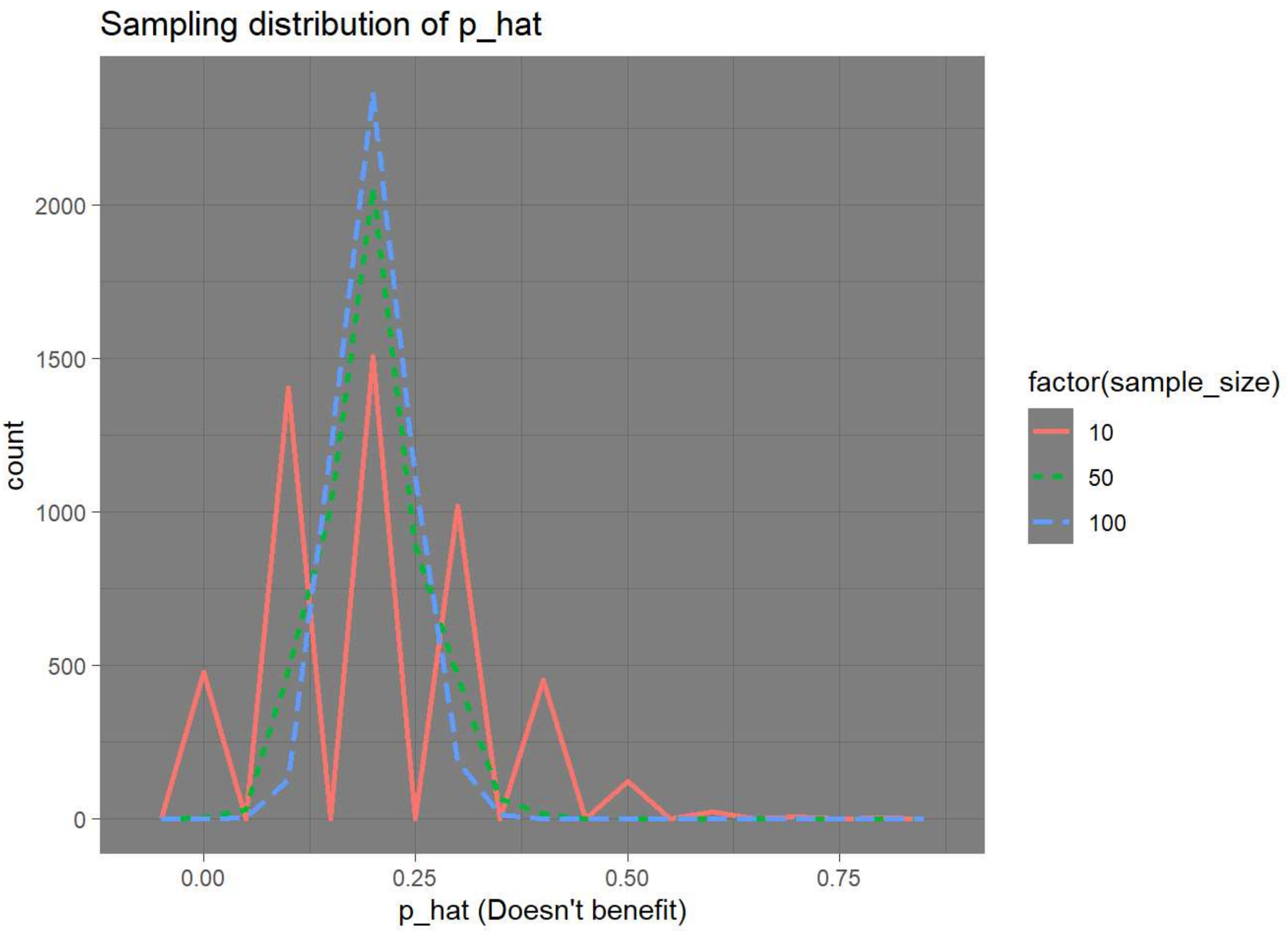
Use the app below to create sampling distributions of proportions of *Doesn’t benefit* from samples of size 10, 50, and 100. Use 5,000 simulations. What does each observation in the sampling distribution represent? How does the mean, standard error, and shape of the sampling distribution change as the 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.)

Response

```r
sample_sizes = c(10,50,100)
sample_props_10_50_100 <- tibble()

for (size in sample_sizes) {
  sample_props <- global_monitor %>%
    rep_sample_n(size = size, reps = 5000, replace = TRUE) %>%
    count(scientist_work) %>%
    mutate(p_hat = n / sum(n)) %>%
    filter(scientist_work == "Benefits") %>%
    mutate(doesnt_benefit = size - n,
           p_hat = 1 - p_hat,
           sample_size = size) %>%
    select(replicate, sample_size, doesnt_benefit, p_hat)
  sample_props_10_50_100 <- bind_rows(sample_props_10_50_100, sample_props)
}

sample_props_10_50_100 %>%
  ggplot(aes(x = p_hat,
             color = factor(sample_size),
             linetype = factor(sample_size))) +
  geom_freqpoly(binwidth = 0.05, linewidth = 1) +
  labs(x = "p_hat (Doesn't benefit)",
       title = "Sampling distribution of p_hat") +
  theme_dark()
```



```
sample_props_10_50_100 %>%
  group_by(sample_size) %>%
  summarize(min = min(p_hat),
            q1 = quantile(p_hat, 0.75),
            median = median(p_hat),
            mean = mean(p_hat),
            q3 = quantile(p_hat, 0.75),
            max = max(p_hat),
            sd = sd(p_hat))
```

```
## # A tibble: 3 × 8
##   sample_size    min    q1 median  mean    q3   max     sd
##         <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1          10 0       0.3     0.2 0.201  0.3   0.8  0.124
## 2          50 0.0200  0.24    0.2 0.200  0.24  0.42 0.0564
## 3         100 0.0700  0.23    0.2 0.201  0.23  0.36 0.0405
```

We see trends emerge consistent with the Central Limit Theorem. As we increase the sample size, we see (i) the mean of the sampling distribution converges to the population mean (though all sizes produce very close estimates), and (ii) the standard deviation decreases. These trends are apparent when plotting the distributions, as well. The larger samples drive tighter distributions with less spread.

Exercise 7

Question

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 enhances their lives?

Response

```
global_monitor %>%
  sample_n(size = 15) %>%
  table() %>%
  data.frame() %>%
  mutate(Proportion = Freq / sum(Freq))
```

```
##    scientist_work Freq Proportion
## 1        Benefits   12        0.8
## 2 Doesn't benefit    3        0.2
```

Despite the limited sample, we obtain a very solid estimate of population proportion at 80%.

Exercise 8

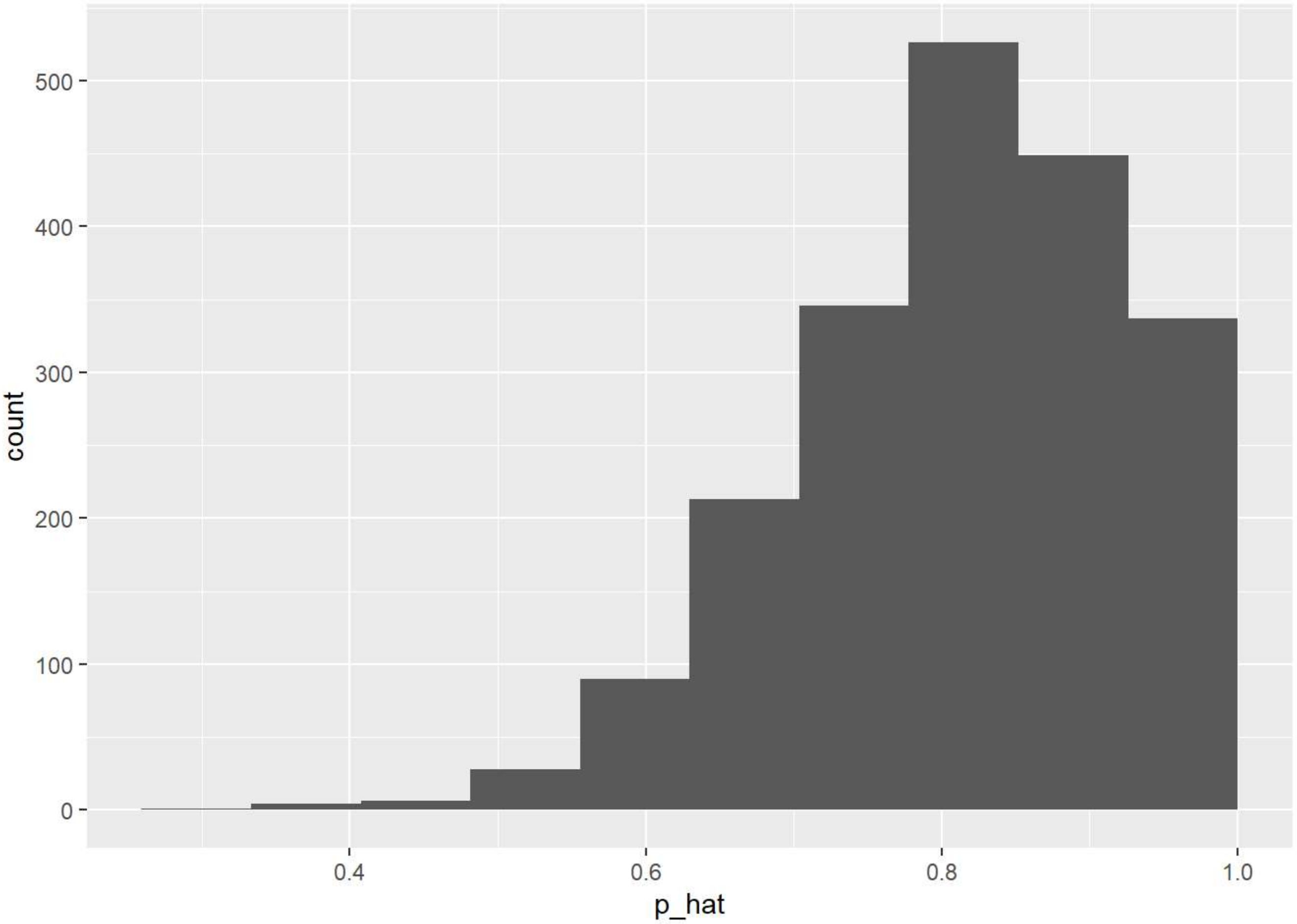
Question

Since you have access to the population, simulate the sampling distribution of proportion of those who think the work scientists do enhances 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 enhances their lives to be? Finally, calculate and report the population proportion.

Response

```
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")

ggplot(data = sample_props15, aes(x = p_hat)) +
  geom_histogram(bins = 10)
```



```
summary(sample_props15$p_hat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3333  0.7333  0.8000  0.7991  0.8667  1.0000
```

The sampling distribution appears left skewed. Because of the low sample size, we see that the CLT does not hold quite as strongly here. However, the sampling distribution also has a mean very close to the population proportion, at 79.9% versus 80%.

Exercise 9

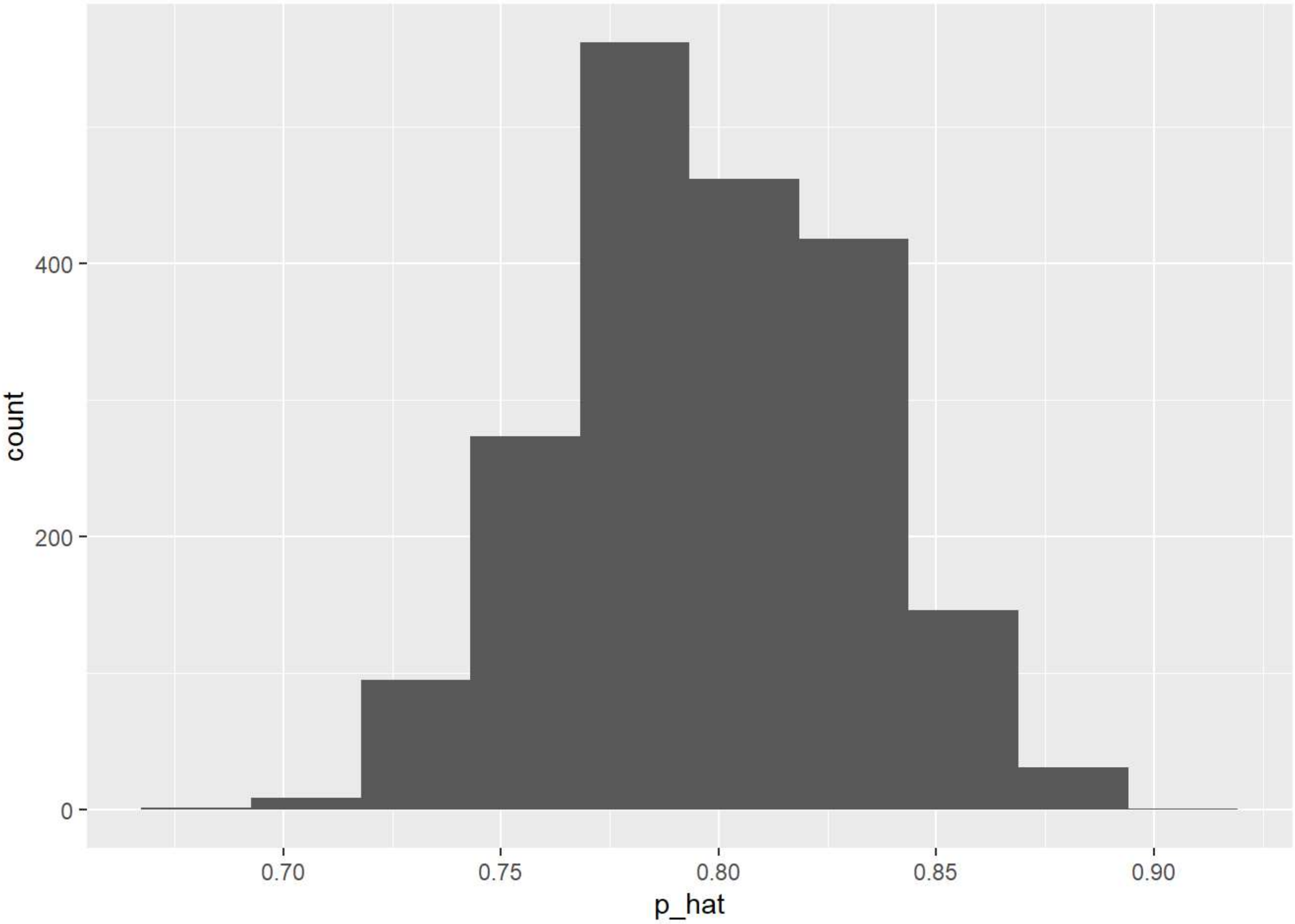
Question

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?

Response

```
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")

ggplot(data = sample_props150, aes(x = p_hat)) +
  geom_histogram(bins = 10)
```



```
summary(sample_props150$p_hat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.6800  0.7733  0.8000  0.7987  0.8200  0.9067
```

With an increased sample size, the sampling distribution appears much more normal. The mean of the sampling distribution, however, remains largely unchanged, and very closely resembles the 80% population proportion.

Exercise 10

Question

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?

Response

Assuming the above question refers to the sampling distributions from 8 and 9, the second distribution from sample size 150 has a smaller spread, as evidenced by the smaller IQR and the tighter visual distribution. A tighter spread is beneficial when making estimates about the population mean, as confidence internvals will also be tighter.