Data 606 Lab5A

Joao De Oliveira

2025-10-09

Overview

In this lab, I will investigate the ways in which the statistics from a random sample of data can serve as point estimates for population parameters. I will be formulating a sampling distribution of my estimate in order to learn about the properties of the estimate, such as its distribution.

Load Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

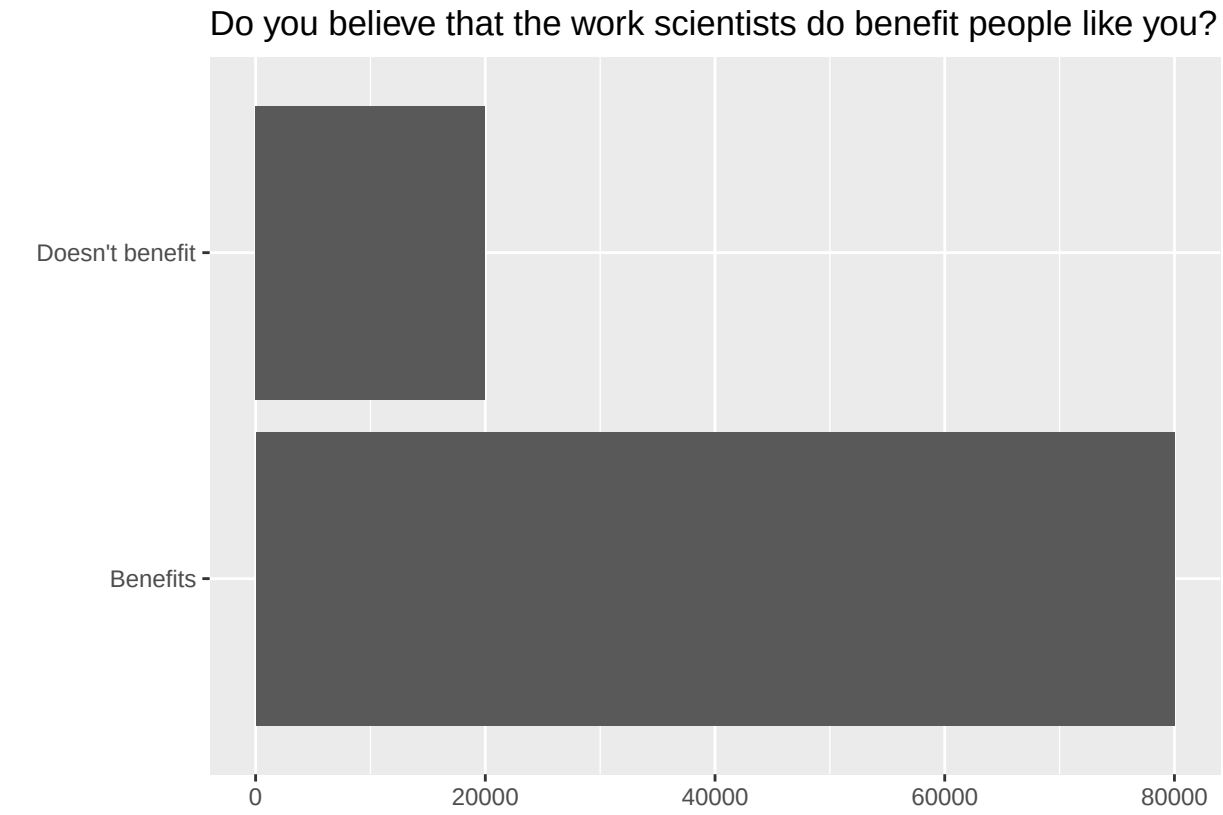
```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
```

Data

```
set.seed(1234)

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

```
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()
```




```
# summary stats
global_monitor %>%
  count(scientist_work) %>%
  mutate(p = n /sum(n))
```

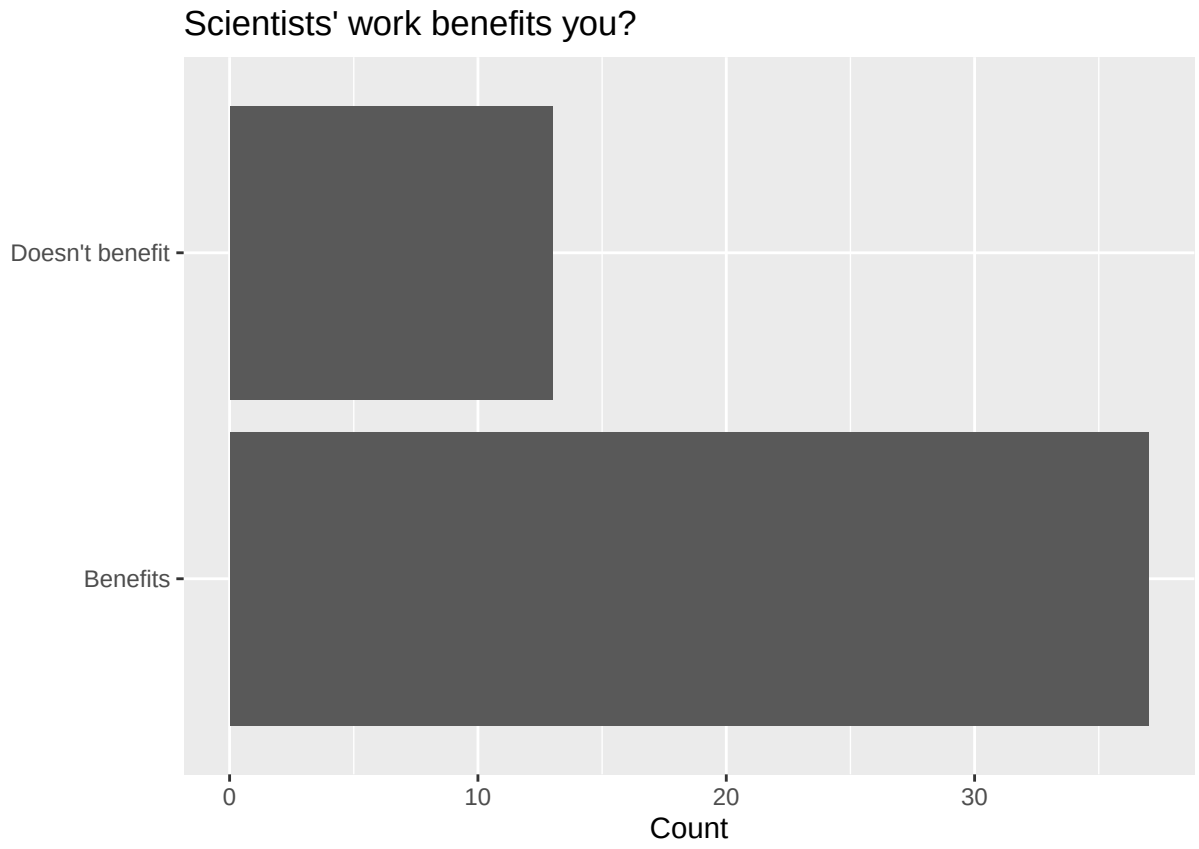
```
## # A tibble: 2 x 3
##   scientist_work      n     p
##   <chr>           <int> <dbl>
## 1 Benefits        80000   0.8
## 2 Doesn't benefit 20000   0.2
```

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

Exercise 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.

```
ggplot(samp1, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "Count",
    title = "Scientists' work benefits you?") +
  coord_flip()
```




```
# sample summary stats
sample_summary <- samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n / sum(n))

sample_summary
```

```
## # 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
```

Answer 1

The distribution of responses in the sample differs slightly from the population distribution. While the population shows an 80/20 split between those who believe the work of scientists benefits them and those who do not, the sample of 50 participants shows a different proportion of approximately 86% versus 14%. This difference is visible in the bar plots, where the bars for the sample are not perfectly aligned with those of the population. Such variation is expected due to random sampling — each sample may slightly over represent or under represent certain possibilities.

Exercise 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.

Answer 2

I would not expect my sample proportion to be exactly the same as another student's sample proportion. Each student's sample is randomly drawn from the same population, so the specific individuals chosen won't be the same. Because of this random variation, the sample proportions (p_hat) will differ. However, I would expect the differences between our sample proportions to be somewhat small, not very large. Since both samples come from the same large population (where the true proportion p = 20), most random samples of size 50 should give sample proportions that are close to 0.20, though not identical. When I compared results, the sample proportions were slightly different but within a few percentage points, confirming that small variation is normal and expected due to random sampling. Different seeds will also give different outcomes.

Exercise 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?

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

p_hat1 <- mean(samp1$scientist_work == "Doesn't benefit")
p_hat2 <- mean(samp2$scientist_work == "Doesn't benefit")

# compare
tibble(
  sample = c("samp1 (n=50)", "samp2 (n=50)"),
  p_hat   = c(p_hat1, p_hat2)
)
```

```
## # A tibble: 2 x 2
##   sample       p_hat
##   <chr>        <dbl>
```

```
## 1 samp1 (n=50)  0.26
## 2 samp2 (n=50)  0.16
```

```r
# Larger samples
samp100 <- global_monitor %>% sample_n(100)
samp1000 <- global_monitor %>% sample_n(1000)

p_hat100 <- mean(samp100$scientist_work == "Doesn't benefit")
p_hat1000 <- mean(samp1000$scientist_work == "Doesn't benefit")

results <- tibble(
  sample = c("samp1 (n=50)", "samp2 (n=50)", "samp3 (n=100)", "samp4 (n=1000)"),
  p_hat  = c(p_hat1, p_hat2, p_hat100, p_hat1000),
  abs_error = abs(p_hat - 0.20)  # deviation from true p = 0.20
)

results
```

```
## # A tibble: 4 x 3
##   sample         p_hat abs_error
##   <chr>          <dbl>     <dbl>
## 1 samp1 (n=50)   0.26    0.06
## 2 samp2 (n=50)   0.16    0.04
## 3 samp3 (n=100)  0.19    0.0100
## 4 samp4 (n=1000) 0.202   0.00200
```

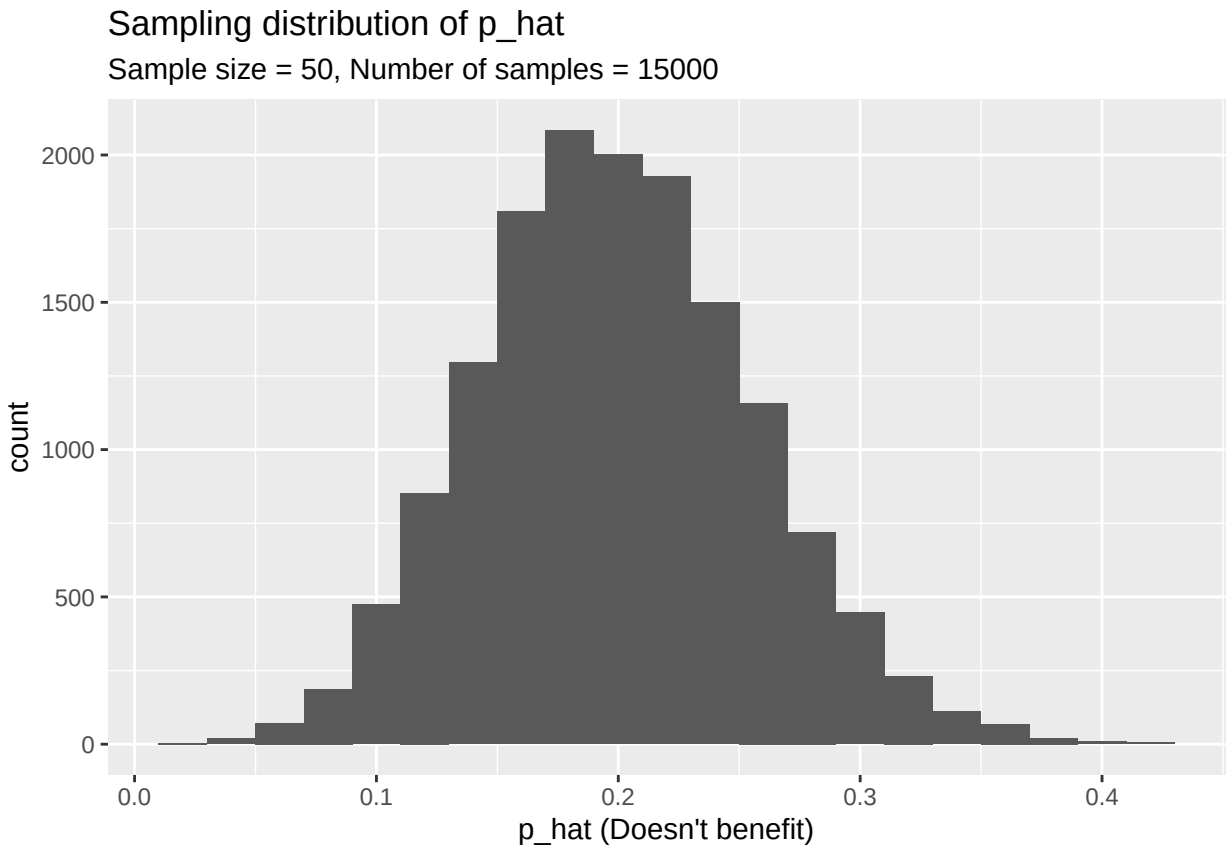
Answer 3

We can see that the sample proportion in sample 1 and sample 2 are different despite both samples having the same size (50). This happens because since the samples are different, there are different representations of the population due to randomness. Samples 3 and 4 will have a better representation of the entire population because they are larger Sample 4, with 1000, is the largest sample size and that also explains why the proportion is pretty close to the population p ($0.186 \sim 0.2$).

Exercise 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.

```r
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")

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"
  )
```




```
# finding num of elements
nrow(sample_props50)
```

```
## [1] 15000
```

Answer 4

sample_props50 has 15000 elements since we are taking one element (the p of not benefiting from scientists work) from each sample of size 50. By looking at the sampling distribution we see that the mean (center) is 0.2, as expected and the distribution has a shape similar to a bell-shaped distribution (normal distribution). The great majority of the values is between 0.1 and 0.3.

Exercise 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?

```
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")
```

```r
# print the data frame
sample_props_small
```

```
## # A tibble: 21 x 4
## # Groups:   replicate [21]
##    replicate scientist_work      n p_hat
##        <int> <chr>           <int> <dbl>
##  1         1 Doesn't benefit     2   0.2
##  2         2 Doesn't benefit     2   0.2
##  3         3 Doesn't benefit     4   0.4
##  4         5 Doesn't benefit     3   0.3
##  5         6 Doesn't benefit     2   0.2
##  6         7 Doesn't benefit     4   0.4
##  7         8 Doesn't benefit     5   0.5
##  8        10 Doesn't benefit     2   0.2
##  9        11 Doesn't benefit     1   0.1
## 10        12 Doesn't benefit     5   0.5
## # i 11 more rows
```

```r
# number of observations
nrow(sample_props_small)
```

```
## [1] 21
```

Answer 5

The number of observations in the 25 reps with size 10 each was 21, which is not what I was expecting (25). However, this is probably due to the fact that some of them will have probability of success ("doesn't benefit) equal to 0, which is not that far from 0.2. So, 4 of the 25 samples got probability equal to zero (everyone in those 4 samples believed that scientist benefit them). So, each observation represent the sample proportion of people who do not believe scientists benefit them (if not zero).

Exercise 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.)

```r
simulate_props <- function(n_size, reps = 5000) {
  global_monitor %>%
    mutate(scientist_work = factor(scientist_work,
                                   levels = c("Benefits", "Doesn't benefit"))) %>%
    rep_sample_n(size = n_size, reps = reps, replace = TRUE) %>%
    count(replicate, scientist_work, name = "count", .drop = FALSE) %>%
    group_by(replicate) %>%
    mutate(p_hat = count / sum(count)) %>%
    filter(scientist_work == "Doesn't benefit") %>%
    ungroup() %>%
```

```
    rename(rep_id = replicate) %>%
    mutate(sample_n = n_size) %>%
    select(rep_id, sample_n, p_hat)
}

# samples with size 10, 50, and 100
sample_props_10 <- simulate_props(10, 5000)
sample_props_50 <- simulate_props(50, 5000)
sample_props_100 <- simulate_props(100, 5000)

# number of observations
nrow(sample_props_10)
```

```
## [1] 5000
```

```
nrow(sample_props_50)
```

```
## [1] 5000
```

```
nrow(sample_props_100)
```

```
## [1] 5000
```

```
# compare empirical mean and sd with theoretical
p_true <- 0.20
all_samples <- bind_rows(sample_props_10, sample_props_50, sample_props_100)

summary_by_n <- all_samples %>%
  group_by(sample_n) %>%
  summarize(
    mean_p_hat = mean(p_hat),
    sd_p_hat = sd(p_hat),
    se_theoretical = sqrt(p_true * (1 - p_true) / unique(sample_n)),
    .groups = "drop"
  )

summary_by_n
```

```
## # A tibble: 3 x 4
##   sample_n mean_p_hat sd_p_hat se_theoretical
##      <dbl>      <dbl>    <dbl>          <dbl>
## 1       10      0.198   0.128          0.126
## 2       50      0.200   0.0558         0.0566
## 3      100      0.200   0.0399         0.04
```

Answer 6

Each observation represents one simulated sample proportion (p hat) calculated from each random sample, which is the proportion who answered "doesn't benefit". We can see that the mean of p_hat is 0.198 (size 10) and 0.20 for both sizes 50 and 100, confirming that p_hat is unbiased for p=0.20 since the mean doesn't

change with sample size. The empirical standard deviation (represented as sd_p_hat) is very close to the theoretical standard deviation and decreases with the size of the sample, indicating low variability for large samples. I increased the number of repetitions (50000 and 100000) to see the change and, as expected, the empirical and theoretical values of standard deviation converge even more than with 5000 repetitions.

Exercise 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?

```
sample_15 <- global_monitor %>%
  sample_n(15)

# proportion "benefits"
p_hat_benefits <- mean(sample_15$scientist_work == "Benefits")
p_hat_benefits
```

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

Answer 7

The sample proportion who believe scientists benefit them is 0.7333 (around 73.3%). This is my point estimate of the population proportion, which is lower than the population proportion of 0.80 (80%). This lower value is explained by the small size of the sample, which makes the results less accurate (vary more).

Exercise 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.

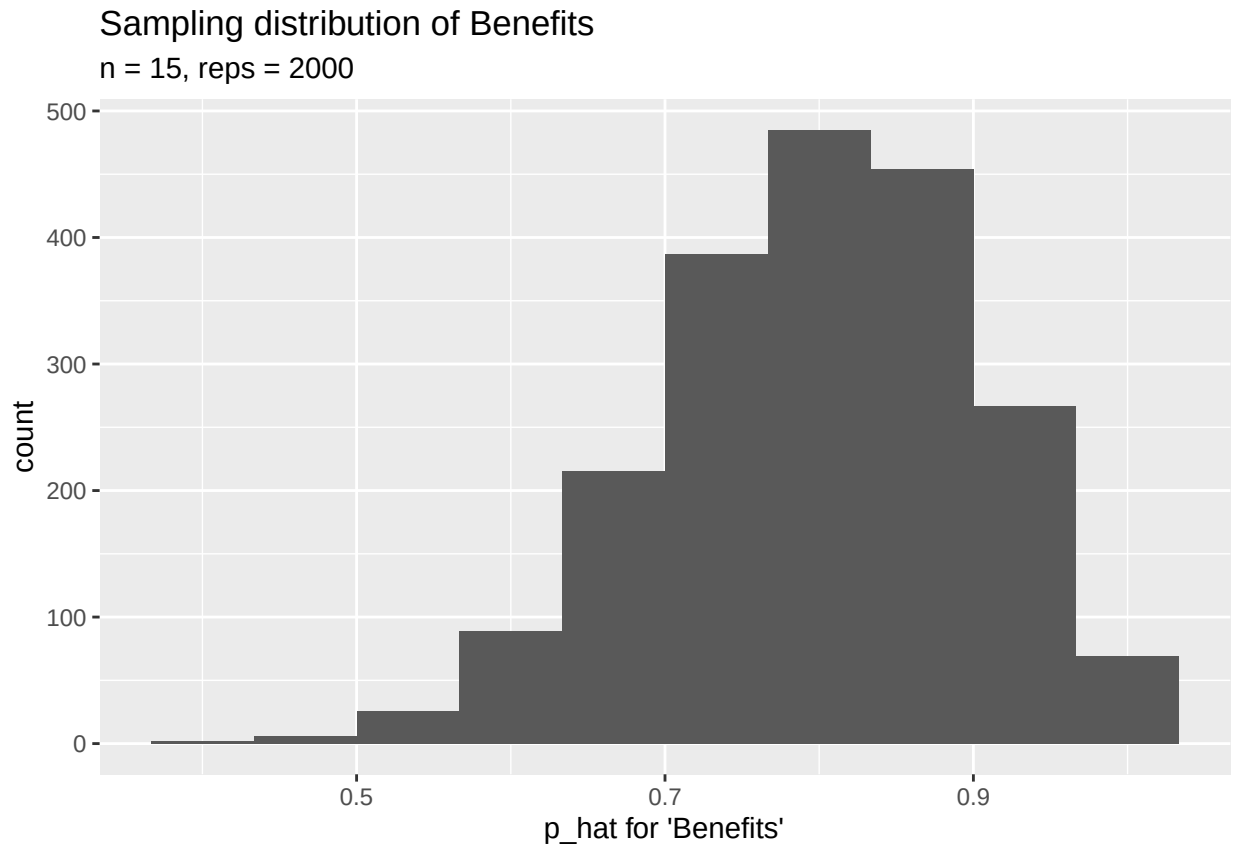
```
global_monitor <- global_monitor %>%
  mutate(scientist_work = factor(scientist_work,
                                 levels = c("Benefits", "Doesn't benefit")))

# sampling distribution size n = 15 and 2000 reps
sample_props15 <- global_monitor %>%
  rep_sample_n(size = 15, reps = 2000, replace = TRUE) %>%
  count(replicate, scientist_work, name = "count", .drop = FALSE) %>%
  group_by(replicate) %>%
  mutate(p_hat = count / sum(count)) %>%
  filter(scientist_work == "Benefits") %>%
  ungroup() %>%
  select(replicate, p_hat)

# number of observations
nrow(sample_props15)
```

```
## [1] 2000
```

```
head(sample_props15)
```

```
## # A tibble: 6 x 2
##   replicate p_hat
##       <int> <dbl>
## 1         1 0.867
## 2         2 0.667
## 3         3 1
## 4         4 0.733
## 5         5 0.533
## 6         6 0.8
```

```
# plot
ggplot(sample_props15, aes(x = p_hat)) +
  geom_histogram(binwidth = 1/15) +
  labs(
    x= "p_hat for 'Benefits'",
    title = "Sampling distribution of Benefits",
    subtitle = "n = 15, reps = 2000"
  )
```



```r
# summary
summ_15 <- sample_props15 %>%
  summarise(
    mean_p_hat = mean(p_hat),
    sd_p_hat = sd(p_hat),
    min_p_hat = min(p_hat),
    max_p_hat = max(p_hat)
  )

summ_15
```

```
## # A tibble: 1 x 4
##   mean_p_hat sd_p_hat min_p_hat max_p_hat
##        <dbl>    <dbl>     <dbl>     <dbl>
## 1      0.799    0.104       0.4         1
```

```r
# theoretical SE
se_theory_15 <- sqrt(0.8 * 0.2 / 15)
se_theory_15
```

```
## [1] 0.1032796
```

```r
# population proportion 'Benefits'
proportion_pop_benefits <- mean(global_monitor$scientist_work == "Benefits")
proportion_pop_benefits
```

```
## [1] 0.8
```

Answer 8

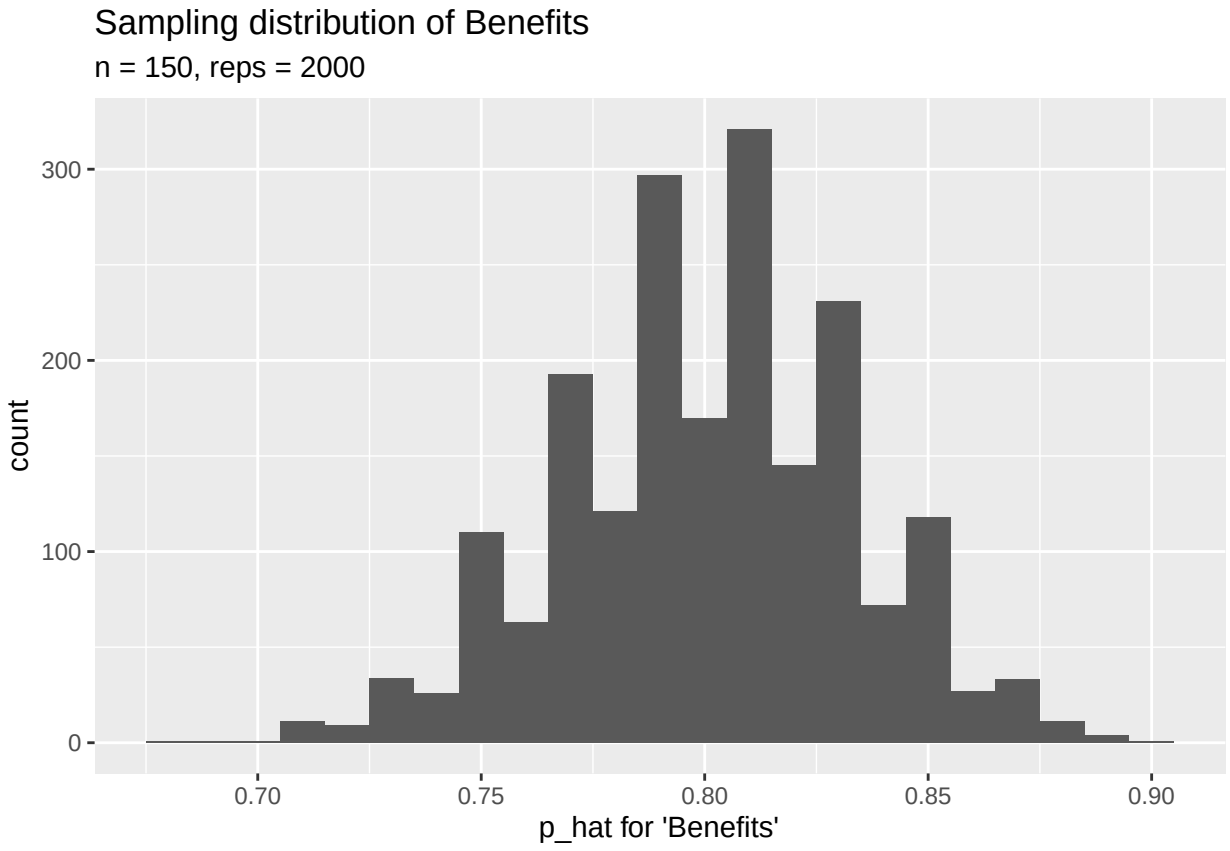
Now each observation represents one sample proportion for "benefits" from a random sample with size 15. By analyzing the graph, we notice an approximate bell-shaped distribution but slightly skewed to the left (longer tail on the left). Based on this sampling distribution the proportion of those who think the work of scientists enhances their lives is 0.799, which is very close to the population proportion of 0.80 as calculated above.

Exercise 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?

```r
# sampling distribution size n = 150 and 2000 reps
sample_props150 <- global_monitor %>%
  mutate(scientist_work = factor(scientist_work,
                                 levels = c("Benefits", "Doesn't benefit"))) %>%
  rep_sample_n(size = 150, reps = 2000, replace = TRUE) %>%
  count(replicate, scientist_work, name = "count", .drop = FALSE) %>%
```

```r
  group_by(replicate) %>%
  mutate(p_hat = count / sum(count)) %>%
  filter(scientist_work == "Benefits") %>%
  ungroup() %>%
  select(replicate, p_hat)

# plot
ggplot(sample_props150, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.01) +
  labs(
    x= "p_hat for 'Benefits'",
    title = "Sampling distribution of Benefits",
    subtitle = "n = 150, reps = 2000"
  )
```




```r
# summary
summ_150 <- sample_props150 %>%
  summarise(
    mean_p_hat = mean(p_hat),
    sd_p_hat = sd(p_hat),
    min_p_hat = min(p_hat),
    max_p_hat = max(p_hat)
  )

summ_150
```

```
## # A tibble: 1 x 4
##   mean_p_hat sd_p_hat min_p_hat max_p_hat
##        <dbl>    <dbl>     <dbl>     <dbl>
## 1      0.801   0.0327      0.68       0.9
```

```
# theoretical SE
se_theory_150 <- sqrt(0.8 * 0.2 / 150)
se_theory_150
```

```
## [1] 0.03265986
```

Answer 9

We can see that the sampling distribution for size of 150 per sample is also approximately bell-shaped, however, is much narrower than the one for size 15. The center is also 0.8, reinforcing the idea that p_hat is unbiased. The theoretical standard error is much smaller since the size increases.Based on this sampling distribution the proportion of those who think the work of scientists enhances their lives is 0.801, which is very close to the population proportion of 0.80 and the guess for a distribution of samples of size 15.

Exercise 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?

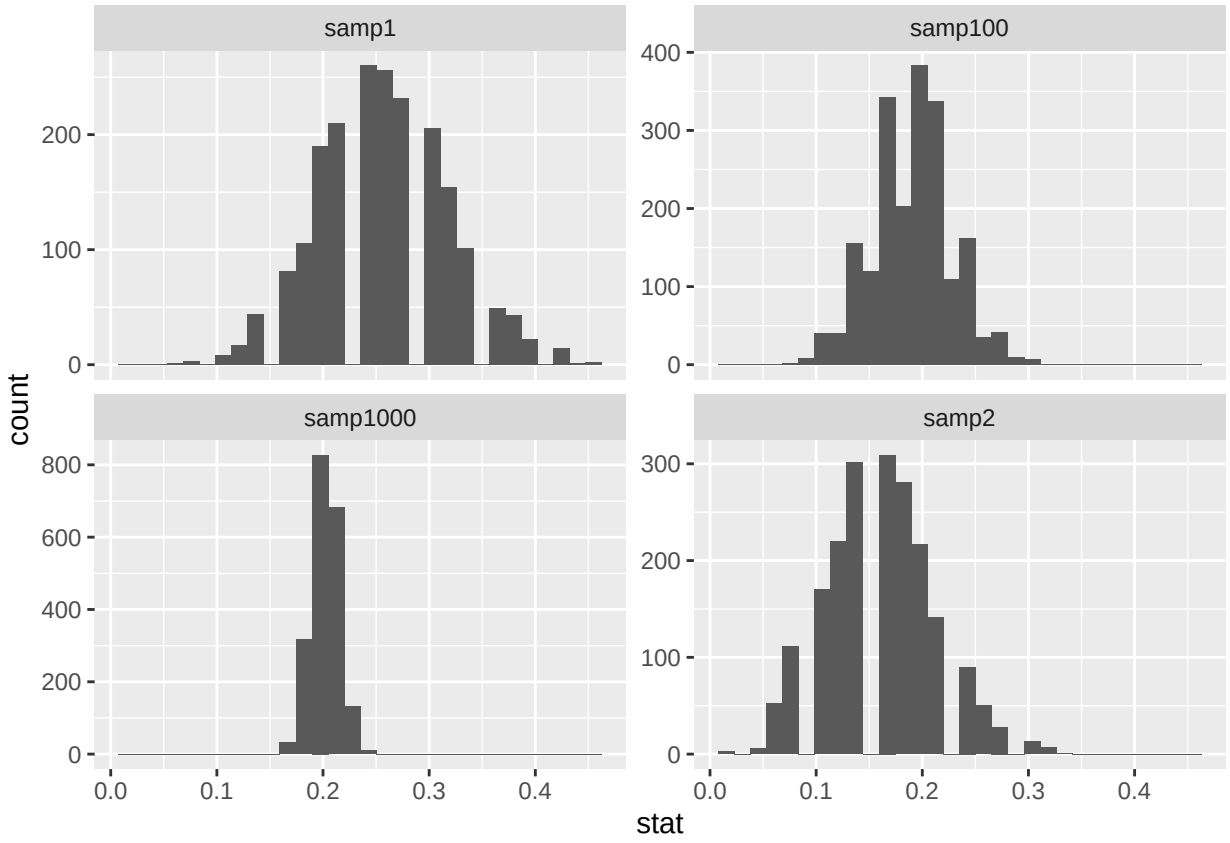
```
bootstrap_props <- function(df, success = "Doesn't benefit", reps = 2000) {
  df %>%
    infer::specify(response = scientist_work, success = success) %>%
    infer::generate(reps = reps, type = "bootstrap") %>%
    infer::calculate(stat = "prop")
}

# using samples from 2 and 3
boot_samp1 <- bootstrap_props(samp1) %>% mutate(sample = "samp1", n = 50)
boot_samp2 <- bootstrap_props(samp2) %>% mutate(sample = "samp2", n = 50)
boot_samp100 <- bootstrap_props(samp100) %>% mutate(sample = "samp100", n = 100)
boot_samp1000 <- bootstrap_props(samp1000) %>% mutate(sample = "samp1000", n = 1000)

boot_all <- bind_rows(boot_samp1, boot_samp2, boot_samp100, boot_samp1000)

# summary
boot_summary <- boot_all %>%
  group_by(sample, n) %>%
  summarise(
    mean_bootstrap = mean(stat),
    sd_bootstrap = sd(stat),
    .groups = "drop"
  ) %>%
  arrange(n, sample)

boot_summary
```

```
## # A tibble: 4 x 4
##   sample       n mean_bootstrap sd_bootstrap
##   <chr>    <dbl>          <dbl>        <dbl>
## 1 samp1       50          0.257       0.0614
## 2 samp2       50          0.161       0.0518
## 3 samp100    100          0.190       0.0391
## 4 samp1000  1000          0.202       0.0128
```

```
# plot
ggplot(boot_all, aes(x = stat)) +
  geom_histogram(bindwidth = 0.01) +
  facet_wrap(~ sample, scales = "free_y")
```

```
## Warning in geom_histogram(bindwidth = 0.01): Ignoring unknown parameters:
## `bindwidth`
```

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



Answer 10

If the question is asking to compare the distributions with 2000 reps for "benefit" with sizes 15 and 150, then the one with size 150 has a much smaller spread as mentioned above (min and max much closer), which makes it a better option, there is more certain in that option. If the question is asking about the distribution samples of 2 and 3, then we can see that the distribution with higher n has a smaller spread. The

2 distributions with size 50 have a much higher spread and a proportion further away from the population proportion than the distributions using larger sample sizes. So, to get estimates closer to the true value, we need smaller spread, which we get with larger sample sizes.