Foundations for statistical inference - Sampling distributions

Alice Ding

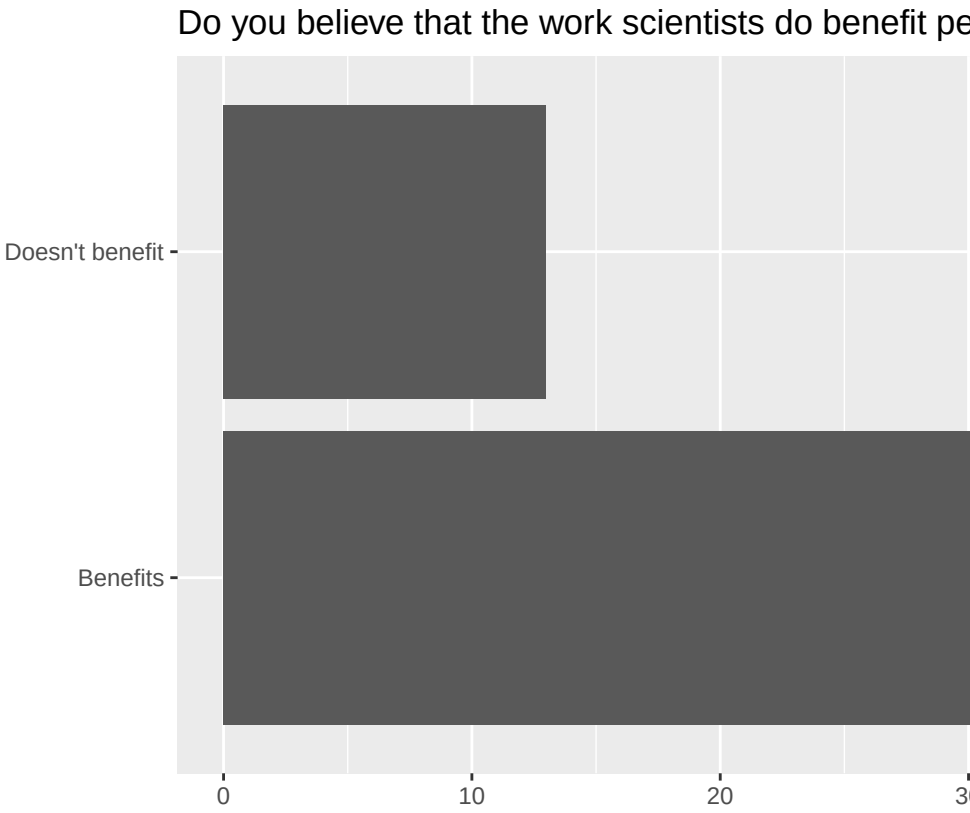
Excercise 1

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

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

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

Describe the distribution of responses in this sample. How does it compare to the distribution



of responses in the population.

```
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
```

Based on this sample, it looks like 13 of the 50 don't believe that the work scientists do benefits them. At ~26%, this is slightly higher than the actual 20% from the population.

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. No, I don't expect my sample to match the one of another student's. The only way we could guarantee the same sample here is if we chose the same seed. The proportions though would be somewhat different, likely not very different. This is due to both our *random* samples likely being representative of the population, albeit not perfect. The person I asked, Nick Climaco, had 6/50 so this confirms ours aren't the same.

Exercise 3

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

samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

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?

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           42  0.84
## 2 Doesn't benefit     8  0.16
```

`samp2` now has 8 (16%) of the sample that doesn't believe that scientists' work benefits the general public. It's a little lower, but closer to the actual 20% of the population.

```
samp100 <- global_monitor %>%
  sample_n(100)

samp100 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           81  0.81
## 2 Doesn't benefit    19  0.19
```

```
samp1000 <- global_monitor %>%
  sample_n(1000)

samp1000 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          798 0.798
## 2 Doesn't benefit   202 0.202
```

The 100 size sample shows a 19% vs. the 20% while the 1000 sample shows a 20.2%. The 1000 sample provides a more accurate sample and this makes sense as the larger the sample, the closer it is to representing the greater population.

Exercise 4

```
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"
  )
```

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 rows in `sample_props50` and the sampling distribution is pretty normal with a center of 0.2. This makes sense as the actual population is 0.2 so the fact that all these samples center around that value is very good.

Exercise 5

```
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")

print(sample_props_small)
```

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?

```
## # 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
## # ... with 11 more rows
```

There are 21 observations in this data frame and each observation represents one sample of 10 with the count of `Doesn't benefit` associated. The fact that there are less than 25 implies that there were 4 observations that didn't have any `Doesn't benefit`, so that's why there are only 21.

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.)

- Size 10: The highest bar is at 0.2, however the data looks skewed more right rather than having a normal distribution. The mean is 0.23 and the SE is 0.11.
- Size 50: Much more normal than size 10, a little skewed right but otherwise pretty normal. The mean is 0.2 and the SE is 0.06. This makes sense in comparison to size 10 as the mean is at 0.2 and there's lower SE.
- Size 100: Less skewness here and much more normal, it trails off more on the left (higher) side of the chart, but otherwise a clear bell curve. The mean is 0.2 and the SE is 0.04, which makes sense as the SE is less but the 0.2 stays consistent when compared to the size 50.

When increasing the number of simulations, the mean, standard error, and shape don't change that much.

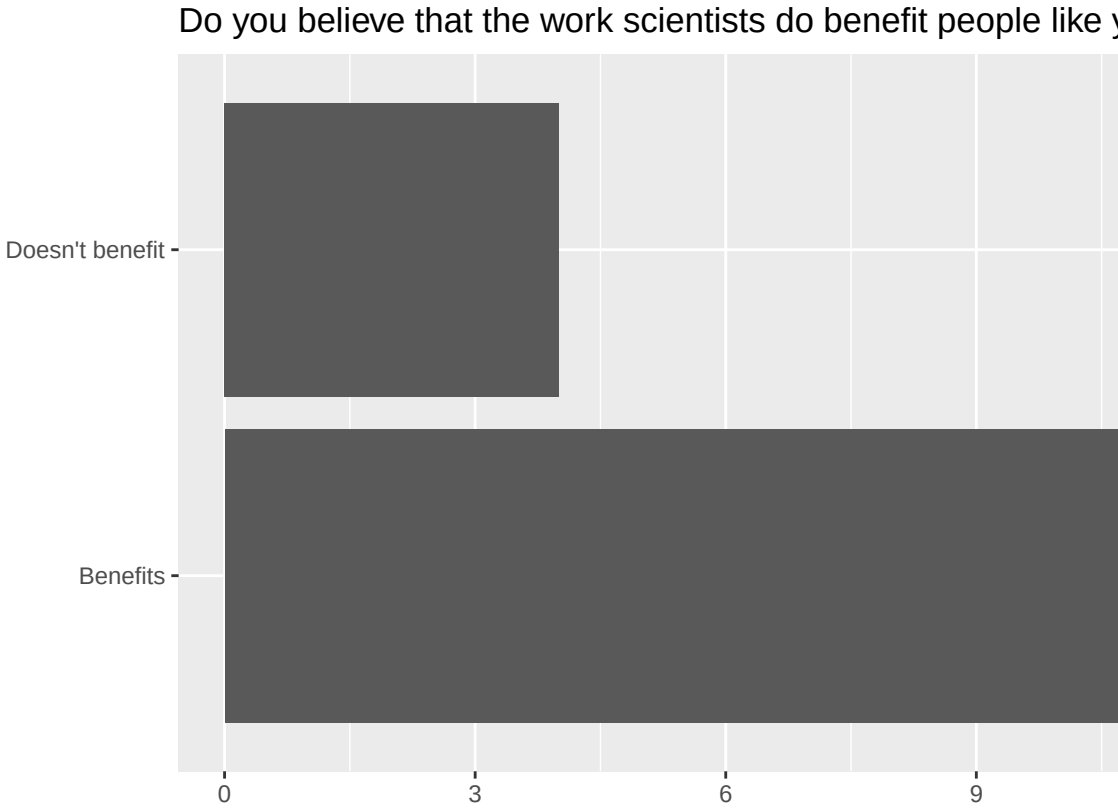
Exercise 7

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

ggplot(samp3, aes(x = scientist_work)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you believe that the work scientists do benefit people like you?"
  ) +
  coord_flip()
```

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 en-



chances their lives?

```
samp3 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n))
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           11 0.733
## 2 Doesn't benefit     4 0.267
```

Based on this sample, it seems like ~73% (11/15) of people believe that scientists' work enhances their lives.

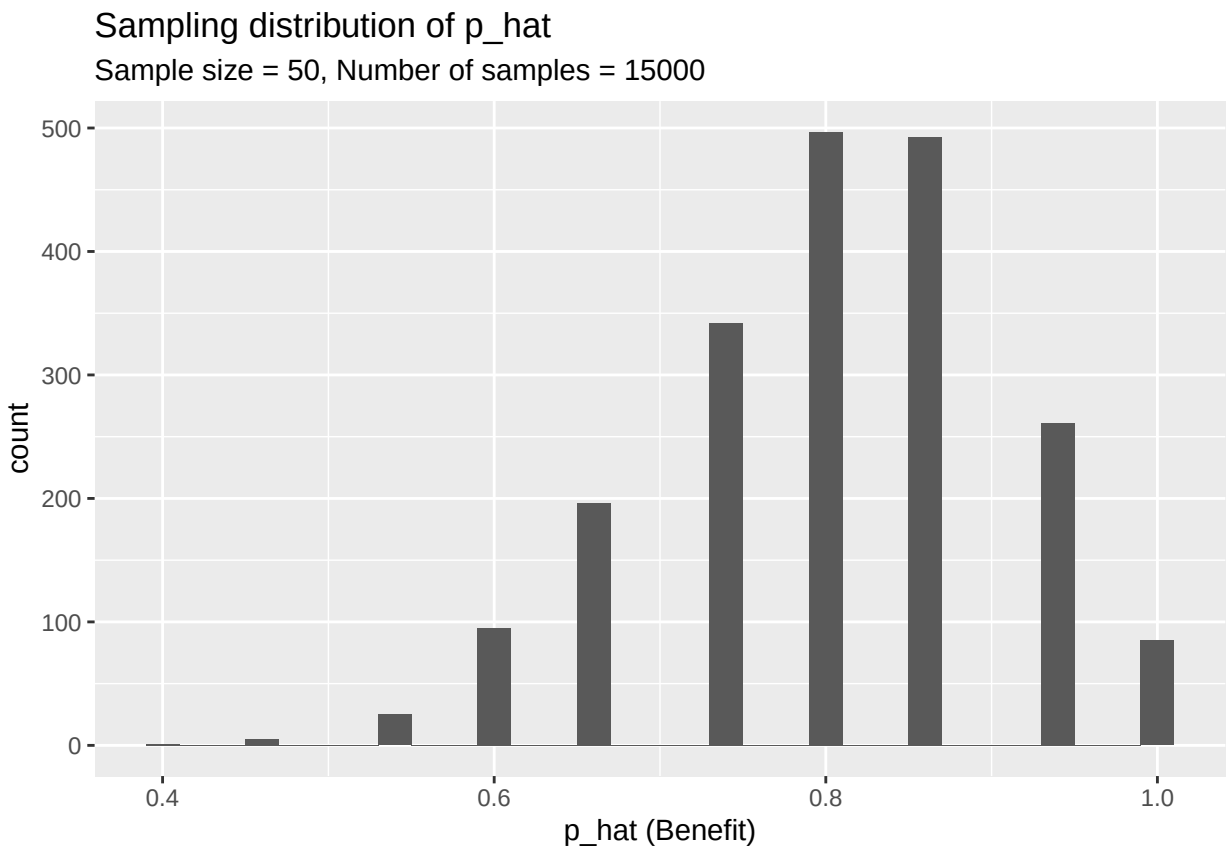
Exercise 8

```
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(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```

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.




```
mean(sample_props15$p_hat)
```

```
## [1] 0.804
```

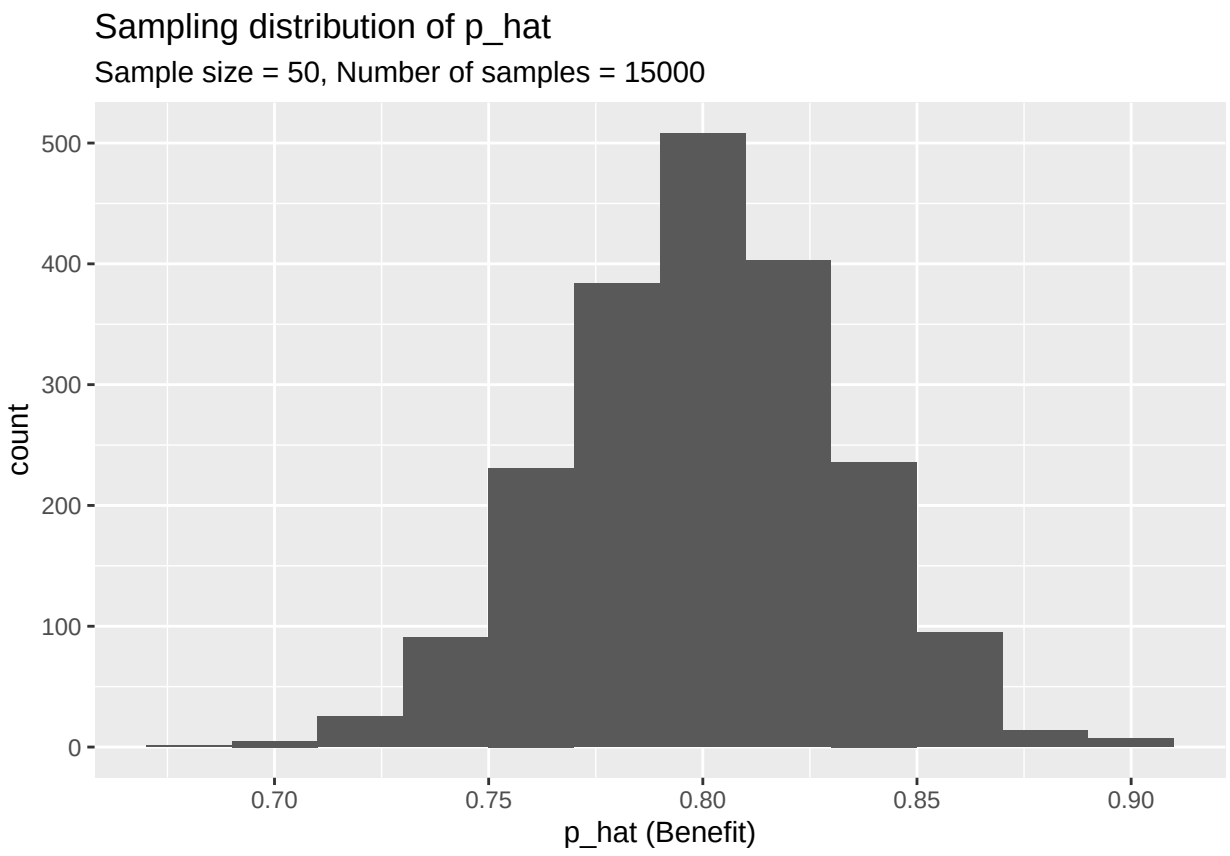
The shape of this distribution looks to be skewed slightly to the left and I would guess that the true proportion would be from 0.8 to 0.85 since that's where the largest columns are. The average of this within the 2,000 observations is 0.804 which makes sense since it hovers around 0.8.

Exercise 9

```
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(binwidth = 0.02) +
  labs(
    x = "p_hat (Benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  )
```

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?




```
mean(sample_props150$p_hat)
```

```
## [1] 0.8000733
```

This data is much more normal in shape with a seemingly perfect bell curve around 0.8. My guess based on this chart is byfar at 0.8 and the average here is 0.8001 which lines up with the guess.

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? I'm assuming we're talking about exercises 8 and 9 here – I would say that the one with the size of 150 has a smaller spread. If we're concerned with making estimates that are more often close to the true value, we'd prefer a sampling distribution with a small spread which means a higher size.