Lab 4 - Sampling Distribution Solution

MATH224 - Intro to Stat

Exercise 1 (4 Points)

1 Point Explanation: Below we see a 84 - 16 split for benefits and doesn't benefit respectively. The sample proportion is 4% off of the true proportion of 80 - 20.

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

samp1 <- global_monitor %>%
  sample_n(50) # 1 Point

samp1 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) # 1 Point
```

```
## # 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
```

Exercise 2 (2 Points)

We can't expect the same proportions for another student unless they are using the same seed or through pure coincidence. We would expect the proportions to vary but not by much. All the proportions would be around the vicinity of the true proportions most of the time.

Exercise 3 (5 Points)

1 Point Explanation: Samp2 seem to have proportions that are much more further away from the true proportions compared to Samp1. Samp2 proportions are 12% away from the true proportions.

1 Point The sampling with sample size of 1000 would provide a more accurate of the population proportion as we know that as sample size increases, the sampling proportion gets closer to the true proportion.

```
samp2 <- global_monitor %>%
  sample_n(50) # 0.5 Point

samp2 %>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) # 0.5 Point
```

```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits           34  0.68
## 2 Doesn't benefit    16  0.32
```

```
global_monitor %>%
  sample_n(100)%>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) # 1 Point
```

```
## # 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
```

```
global_monitor %>%
  sample_n(1000)%>%
  count(scientist_work) %>%
  mutate(p_hat = n /sum(n)) # 1 Point
```

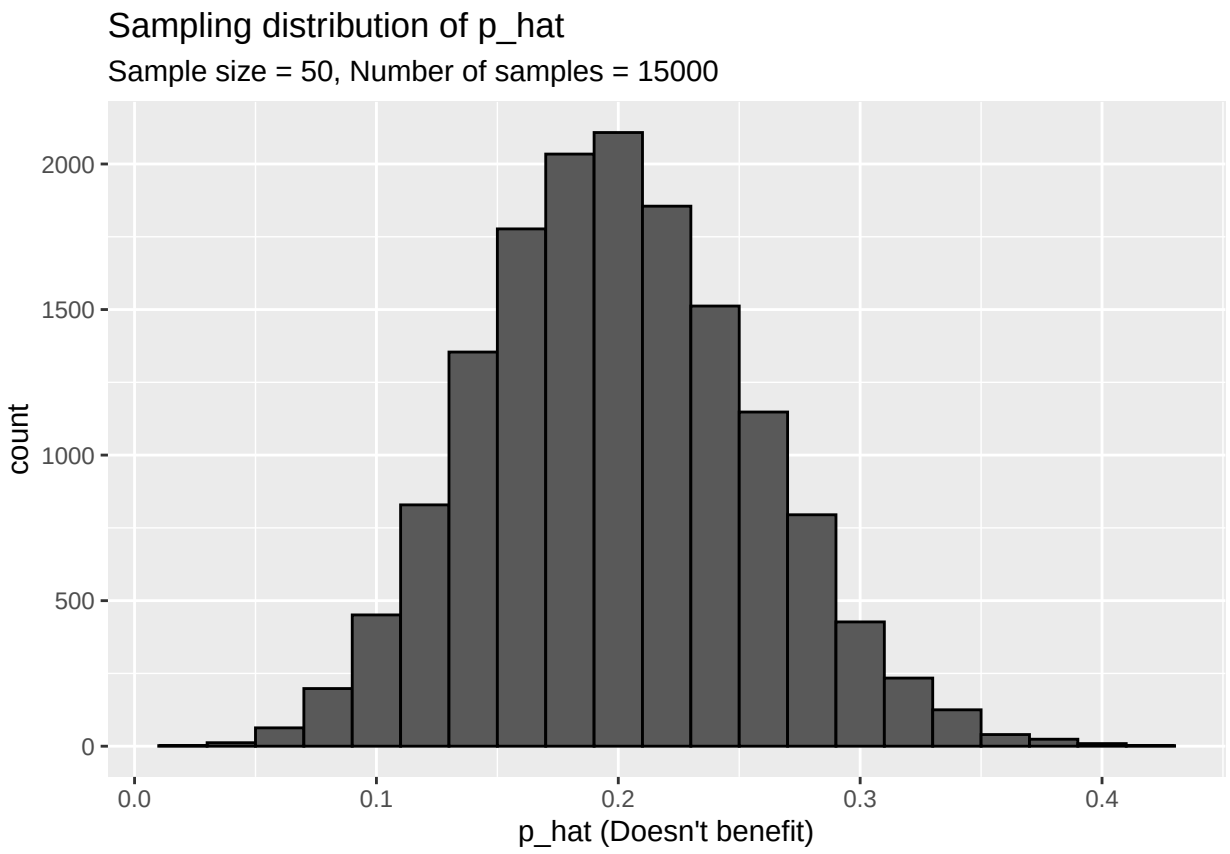
```
## # A tibble: 2 x 3
##   scientist_work      n p_hat
##   <chr>           <int> <dbl>
## 1 Benefits          811 0.811
## 2 Doesn't benefit   189 0.189
```

Exercise 4 (4 Points)

1 Point `sample_props50` has 15,000 observations (elements). The sampling distributions seems to be normal with the center at 0.2, the true proportion for Doesn't Benefit.

```
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") # 1.5 Points

ggplot(data = sample_props50, aes(x = p_hat)) +
  geom_histogram(binwidth = 0.02, col = "black") +
  labs(
    x = "p_hat (Doesn't benefit)",
    title = "Sampling distribution of p_hat",
    subtitle = "Sample size = 50, Number of samples = 15000"
  ) # 1.5 Points
```



Exercise 5 (5 Points)

Link to the app

What does each observation in the sampling distribution represent?

1 Point Each observation in the sampling distribution (x - axis) represents the phat value for the number of simulations we ran.

How does the mean, standard error, and shape of the sampling distribution change as the sample size increases?

As sample size increases,

- Mean gets closer to 0.2 (Population/True proportion). **1 Point**
- Standard Error (SE) decreases. **1 Point**
- Shape of the sampling distribution gets closer and closer to being normal. **1 Point**

How (if at all) do these values change if you increase the number of simulations?

1 Point Increasing the number of simulation makes the sampling distribution more normal. After a certain number of simulations, the mean and SE don't change much at all for sample sizes of 50 and 100