Stats 13 Lab 3

Author Name

2024-08-22

```
# Load the tidyverse, openintro and infer packages, or libraries.

library(tidyverse)
library(openintro)
library(infer)
# If any of these do not work since you do not have them installed yet, run
# install.packages('infer') in console
# install.packages(infer) -> installs 'infer' package on pc forever
# library(infer) -> tells file you want to use packages from the 'infer' package
# (replace infer with other package name if needed)

# Setting a seed -> makes it such that every time you call a random
# function, it returns the same result (helpful for testing purposes)
set.seed(42)
sample.int(n = 100, size = 1) # 49
```

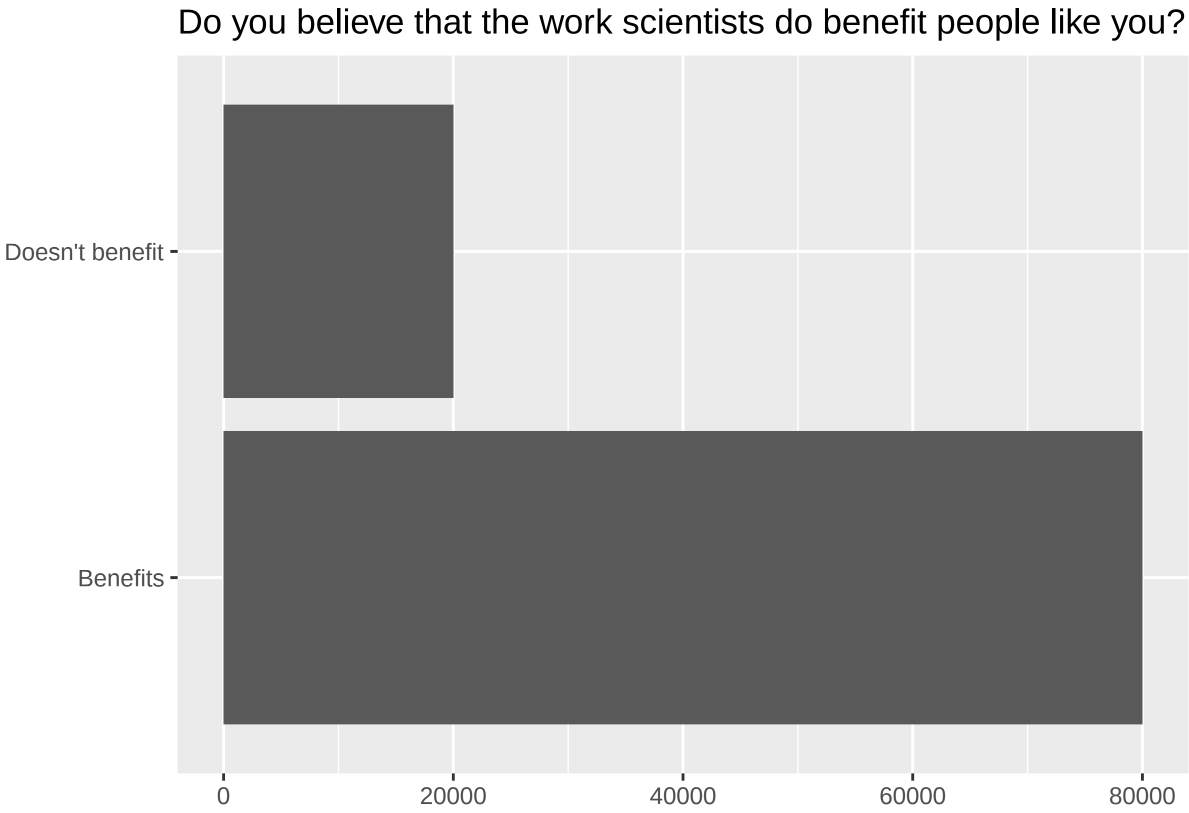
```
## [1] 49
```

```
set.seed(42)
sample.int(n = 100, size = 1) # same result
```

```
## [1] 49
```

Exercise 1

```
# Create a dataframe/tibble that reports 100,000
# responses to the question:
# "Do you believe that the work scientists do benefit people like you?"
global_monitor <- tibble(
  scientist_work = c(rep("Benefits", 80000), rep("Doesn't benefit", 20000))
)

# Create a bar plot of global_monitor
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()
```



```r
# Calculate the proportion of each response
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
```

```r
# Set the seed to 42
set.seed(42)

# Sample 50 responses
samp1 <- global_monitor %>%
  sample_n(50)

### Create a bar plot of your sample


### Calculate the proportion of each response in your sample
```

Exercise 2

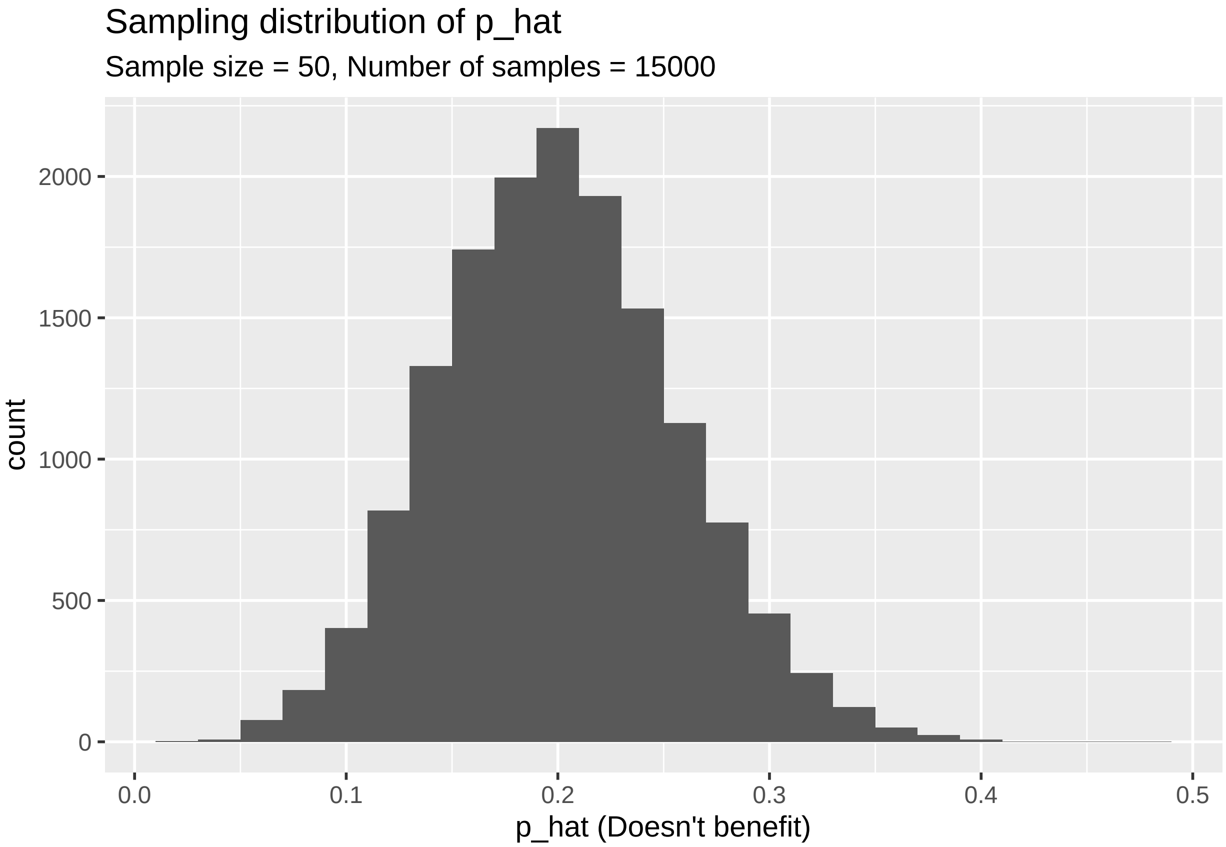
```
### Would the bar plots match if you were to change the seed to a different
### number and take another sample?


### Would the proportions be similar if you were to change the seed to a
### different number and take another sample?
```

Exercise 3

```
### Set the seed to 0


### Sample another 50 responses as samp2


### Calculate the proportion of each response in samp2


### Sample another 100 responses as samp3


### Calculate the proportion of each response in samp3


### Sample another 1000 responses as samp4


### Calculate the proportion of each response in samp4


### What do you notice about the proportions as the sample size increases?
### Will this always be true if you take samples of size 50, 100, 1000?
```

Exercise 4

```
# Obtain 15000 samples of size 50 and calculate the proportion of
# "Doesn't benefit" in each sample
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")

# Create a histogram of your 15000 p_hats
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"
  )
```



```
### Obtain 15000 samples of size 100 and calculate the proportion of
### "Doesn't benefit" in each sample


### Create a histogram of your 15000 p_hats


### How are the two histograms different?
```

Exercise 5

```
### Obtain 25 samples of size 10 and calculate the proportion of
### "Doesn't benefit" in each sample


### Create a histogram of your 25 p_hats


### Why does this histogram look so different from the ones in Exercise 4 and 5?
```

```
# Knit (or generate) the R Markdown file and submit as your TA instructs.
```