Problem set 5

Vivian Yeh

Part I: Confidence intervals

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(stat20data)
data(flights)
```

Question 1

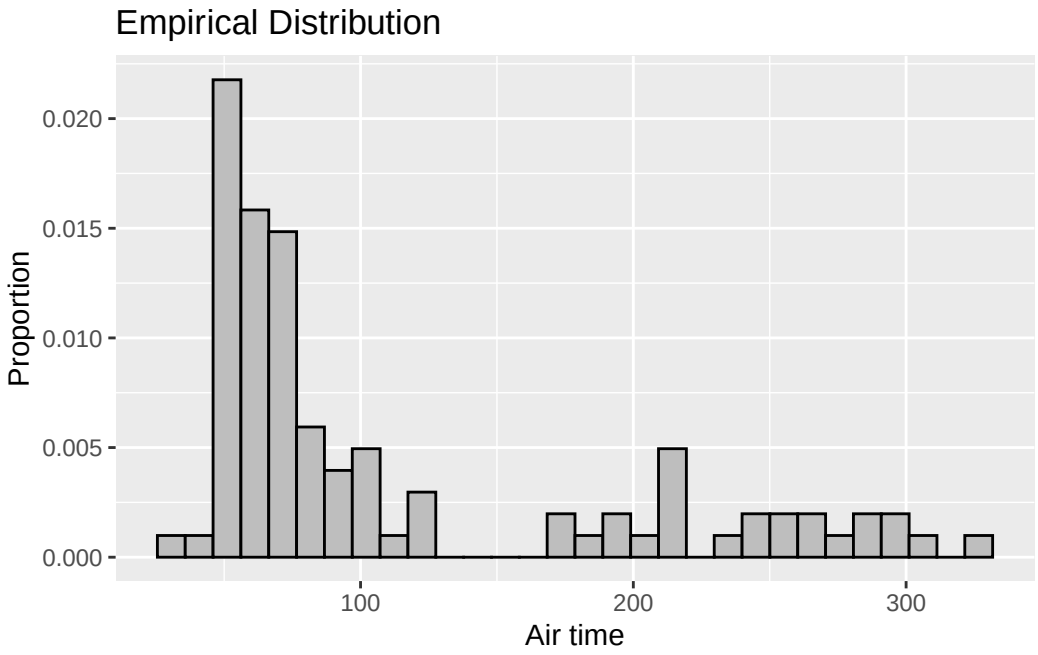
Set a seed of your choosing. Take a random sample of 100 flights from flights. Plot the empirical distribution of air time.

```
set.seed(805)
samp_air_time <- sample(flights$air_time, 100)

emp_dist <- ggplot(as.data.frame(samp_air_time), aes(x=samp_air_time, y=..density..)) +
      geom_histogram(color="black", fill="gray") +
      xlab("Air time") +
      ylab("Proportion") +
```

```
      ggtitle("Empirical Distribution")

emp_dist
```

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

```
Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Question 2

Calculate the mean, standard deviation, and standard error of air time in your sample

```
summarize(as.data.frame(samp_air_time), mean_arr = mean(samp_air_time, na.rm=TRUE), sd_arr
```

```
  mean_arr   sd_arr sd_error
1 113.0707 80.65715 8.065715
```

For the air time in my sample, the mean is 113.0707, standard deviation is 80.65715, and the standard error is 8.05715.

Question 3

Assuming the the sampling distribution is roughly normal, what is your 68% confidence interval? Calculate it using the sample statistics form the previous question.

My 68% confidence interval is 32.41355 to 193.72785.

Next calculate the population mean i.e. the average air time of the entire flights data set. Did your interval contain the population mean?

```
flights %>%
  summarize(mean(air_time, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `mean(air_time, na.rm = TRUE)`
                           <dbl>
1                           129.
```

The population mean is 128.5674, and the population is contained in my interval.

Part 2: Bootstrapping

Question 4

a. The median is 0.3 because the half of 1000 is 500, and the 500th is located at 0.3 interval. The statistic is the sample proportion, which is 0.289, and the parameter is the population proportion, which is unknown.

b. The center of the histogram should lie on 0.289.

c. Using the histogram, I estimate a 90% confidence interval that between 0.23 and 0.35 for the proportion of YouTube videos which take place outdoors.

d. We can be 95% confident that between 22% and 35% of all YouTube videos take place outdoors.

Question 5

Bootstrap distributions of $\hat{}$ Each of the following four distributions was created using a different dataset. Each dataset was based on $= 23$ observations.

i. $= 0.05$: A or D
ii. $= 0.25$: A, B, C, or D
iii. $= 0.45$: B or C
iv. $= 0.55$: B
v. $= 0.75$: None