MATH 254 - Statistical Modeling and Applications - Lab 1

Sampling distributions, Central Limit Theorem

Tural Sadigov

9/5/22

NAME: _____

Who are you working with? _____

PART 1 - Sampling distribution/Central Limit Theorem

Consider the data regarding departure delays from Syracuse Airport in 2019 by three major airlines: United Airlines, Delta Airlines and American Airlines. Data is downloaded from Bureau of Transportation Statistics and edited by cleaning few rows at the top of the CSV file. The data is available to you on GitHub. One can directly read data from GitHub to RStudio using the url of the raw file. See the code below for reading the csv file, and displaying the first 5 rows.

```
library(tidyverse)
delays <- read_csv(file = "https://raw.githubusercontent.com/turalsadigov/MATH_254/main/da
delays %>%
  head(5)
```

```
# A tibble: 5 x 6
  Carrier Date      `Flight #` `Tail #` Destination `Departure Delay Minutes`
  <chr>   <chr>          <dbl> <chr>    <chr>                             <dbl>
1 UA      1/8/2019         714 N810UA   ORD                                  32
2 UA      1/9/2019         714 N4888U   ORD                                  -2
3 UA      1/10/2019        714 N816UA   ORD                                   0
4 UA      1/11/2019        714 N4888U   ORD                                  28
5 UA      1/12/2019        714 N897UA   ORD                                  -1
```

Understanding the data at hand.

1. What does each row represent? Also, characterize each variable (except date) in the data as discrete numerical variable, continuous numerical variable, regular categorical variable or ordinal categorical variable. Code below will help you glimpse into data. Turn on the evaluation.

```
delays %>%
  glimpse()
```

Underlying population.

2. Display the distribution of "Departure delay (Minutes)" (this is the underlying population) in a histogram (with density, not frequency) along with smoothed histogram using/filling code below. Make sure you have titles, labels etc. Insert your name into the title. Find the mean and standard deviation of the distribution. Interpret the distribution, i.e., comment on it. Turn on the evaluation while rendering.

```
# distributions
delays %>%
  ggplot(aes(x = ...,
             y = ..density..)) +
  geom_...(bins = 100,
         color = 'white',
         fill = 'darkgreen') +
  geom_...(lwd = 3,
           col = 'red') +
  ggtitle(...) +
  xlab(...) +
  ylab(...)


# numerical summaries
delays %>%
  summarise(mu = mean(...),
            sigma = ...(...))
```

Sampling distribution of sample Inter Quartile Range (IQR)

3. Construct the simulated sampling distribution of sample IQR of delay minutes of 20 flights with 2000 samples. Impose the smoothed density on it. Comment on the sampling distribution (note that there are $3.8389 * (10^{51})$ possible samples of 20 for this data, and we are only looking at 2000 of such samples). Turn on the evaluation while rendering.

```
library(infer)
set.seed(2022)
delays %>%
  rep_sample_n(size = 20,
               reps = 2000,
               replace = FALSE) %>%
  group_by(replicate) %>%
  summarise(iqr = IQR(`Departure Delay Minutes`)) %>%
  ggplot(aes(x = ... , y = ..density..)) +
  geom_...(...) +
  geom_...(...)
```

4. Find the mean and standard error of the sampling distribution of the sample IQR. Turn on the evaluation while rendering.

```
set.seed(2022)
delays %>%
  rep_sample_n(size = ...,
               reps = ...,
               replace = FALSE) %>%
  group_by(...) %>%
  summarise(iqr = ...) %>%
  summarise(average = mean(...),
            standard_error = sd(...))
```

5. Can sampling distribution of the sample IQR be considered as approximately normal? Does CLT apply? Explain.

6. Construct the simulated sampling distribution of sample IQR of delay minutes of 200 flights with 2000 samples. Impose the smoothed density on it. Comment on the sampling distribution. Can sampling distribution of the sample IQR be considered as approximately normal? Does CLT apply? Turn on the evaluation while rendering.

```
set.seed(2022)
...
```

7. Compare sampling distribution in (3) and (6) above. What are the striking characteristics that make two sampling distribution different?

PART 2 (Independent from above): Parameter of interest exercises

In the following situations, describe a potential parameter of interest of the given study, in the context of the problem as well as its symbol (such as $\mu, \sigma, ...$ etc.), and also mention the sample statistics that approximate the parameter.

8. In a random sample of 765 adults in the United States, 322 say they could not cover an unexpected $400 expense without borrowing money or going into debt.
   - What could be the parameter of interest of the study?
   - What is the sample statistic that approximates it?
9. The nutrition label on a bag of potato chips says that one ounce (28 gram) serving of potato chips has 130 calories and contains 10 grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded sample mean of 134 calories with a standard deviation of 17 calories.
   - What could be the parameter of interest of the study?
   - What is the sample statistic that approximates it?
10. A remote-control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection management says they will switch suppliers, otherwise they will maintain the current supplier. Quality control engineer from collects sample of gears, examining 1000 gears from each company and finds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier.
    - What could be the parameter of interest of the study?
    - What is the sample statistic that approximates it?