R Assignment 03

Digvijay Jondhale 0862899

Question 1- Shapiro-Wilks Unmasked [12 marks]

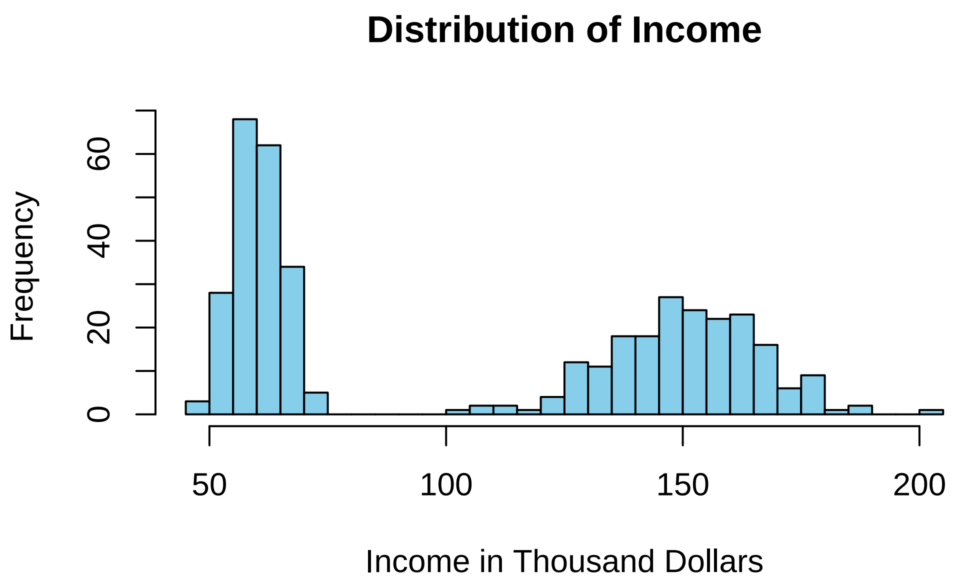
Loading the data ...

```
income <- readRDS("income.rds")
income_df <- data.frame(income)
```

a) [4 marks] Plot a histogram, QQ-plot, and boxplot of the data. Also run shapiro.test. Would you conclude that the normality condition is satisfied?

Ans : ### Histogram

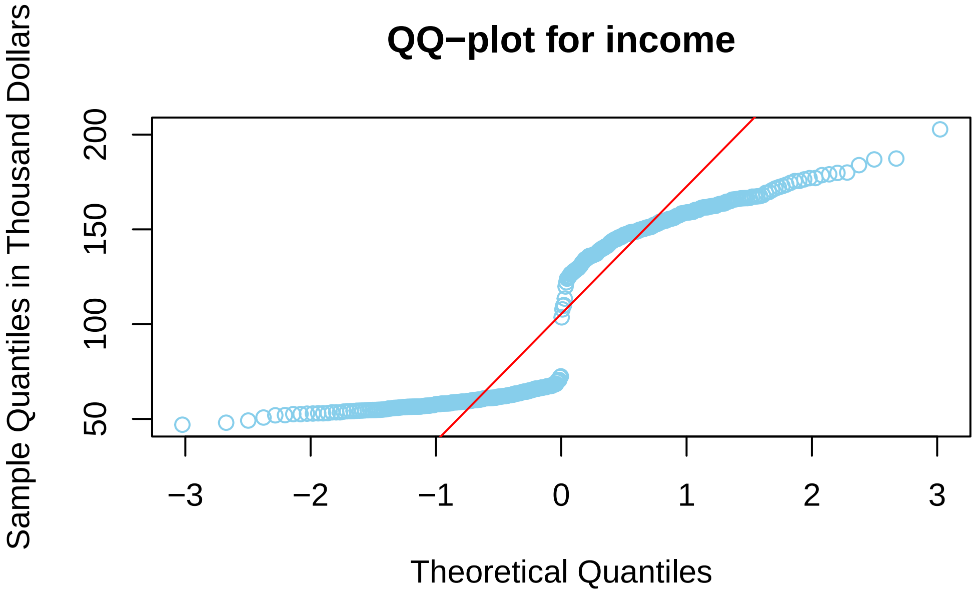
```
hist(income_df$income/1000,breaks = 30,
     xlab = "Income in Thousand Dollars",
     ylab = "Frequency",
     col = "skyblue",
     main = "Distribution of Income"
     )
```



Bimodal , Not Normal !

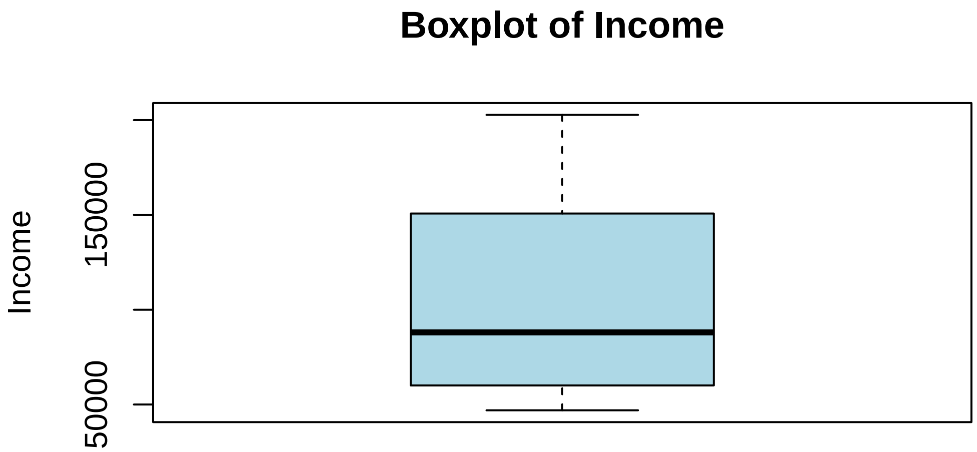
QQ-plot

```
# QQ-Plot
qqnorm(income_df$income/1000,
       main = "QQ-plot for income",
       ylab = "Sample Quantiles in Thousand Dollars",
       col = "skyblue")
qqline(income_df$income/1000, col = "red")
```



The points deviate from the reference line , not Normal !

```
# Boxplot
boxplot(income_df$income,
        main = "Boxplot of Income",
        ylab = "Income",
        col = "lightblue")
```



- the median is not at the center of box , indicating that the data is not normally distributed.
- The whisker lengths are different , indicating skewness in data, hence not Normal.

```
shapiro.test(income_df$income)
```

```

	Shapiro-Wilk normality test

data:  income_df$income
W = 0.8115, p-value < 2.2e-16
```

As the p-value is < 0.05 , we can say that the data of income is not normally distributed

Results from graph summarized :

- Histogram : Bi-modal , Not Normally Distributed.
- QQ-Plot : Deviation from the diagonal line, Not Normally Distributed.
- Box-Plot : asymmetric whisker lengths and median not at center, Not Normally Distributed.
- Shapiro-Wilk normality test : p-value < 0.05 , Not Normally Distributed.

All the results indicates a strong evidence against the normality condition.

b) [4 marks] Use the infer package to generate 4000 bootstrap samples of the sample mean and plot a histogram, QQ-plot, and boxplot of these bootstrap sample means.

Ans :

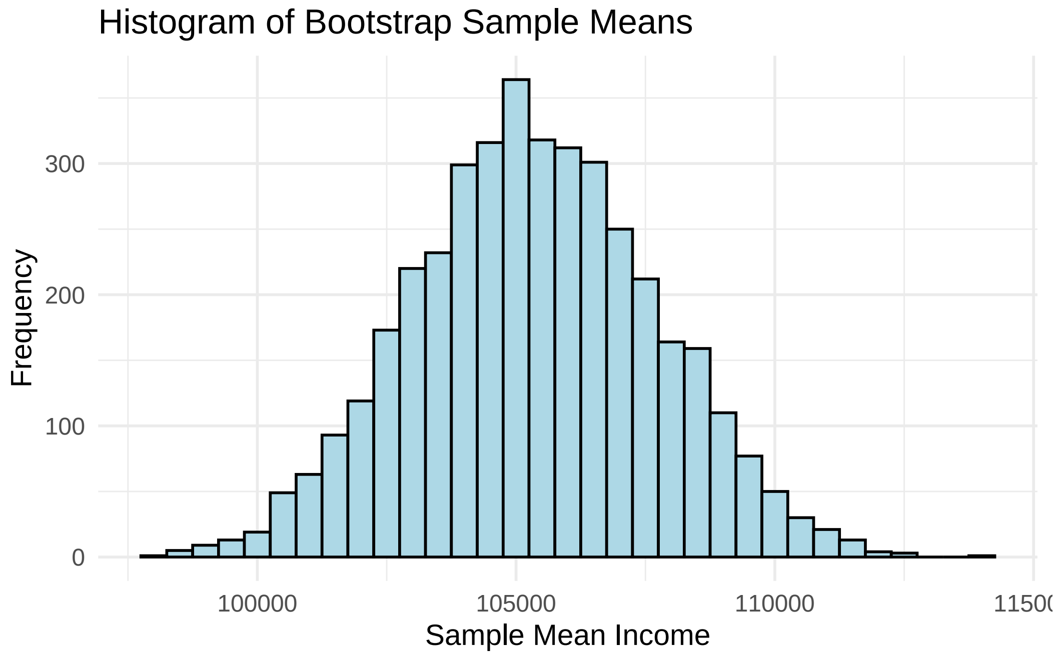
```
library(infer)
library(ggplot2)
```

boot-straping 4000 sample means

```
set.seed(0862899) # Student number
bootstrap_samples <- income_df %>%
  specify(response = income) %>%
  generate(reps = 4000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

Histogram of bootstrap means

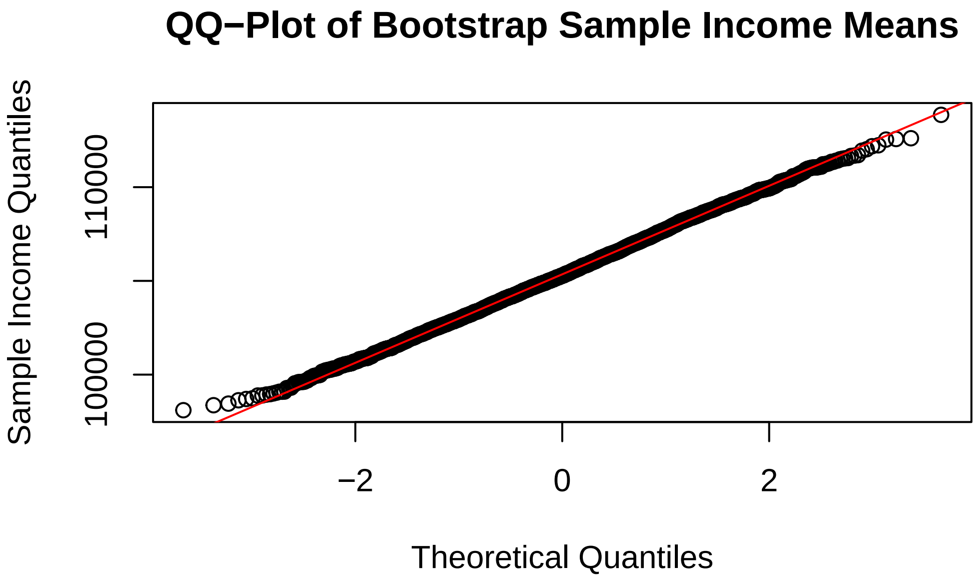
```
ggplot(bootstrap_samples, aes(x = stat)) +
  geom_histogram(binwidth = 500 ,fill = "lightblue", color = "black") +
  labs(title = "Histogram of Bootstrap Sample Means", x = "Sample Mean Income", y = "Frequen
  theme_minimal()
```



- Approximately Normal

QQ-Plot of bootstrap means

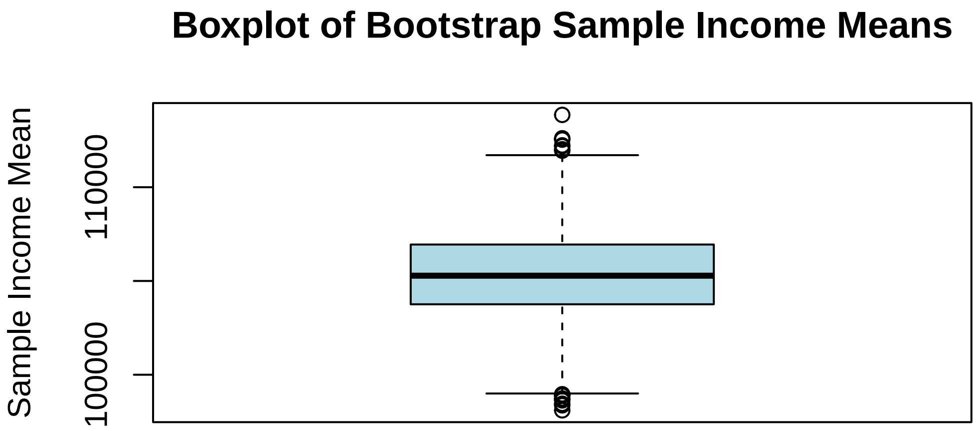
```
# Plot QQ-Plot of the bootstrap sample means
qqnorm(bootstrap_samples$stat, main = "QQ-Plot of Bootstrap Sample Income Means",ylab = "Samp
qqline(bootstrap_samples$stat, col = "red")
```



- It does not deviates from the diagonal line , means the sample bootstrap data is approximately normally distributed.

Box-Plot of bootstrap means

```r
boxplot(bootstrap_samples$stat,
        main = "Boxplot of Bootstrap Sample Income Means",
        ylab = "Sample Income Mean",
        col = "lightblue")
```



- The median is exactly at center and the whisker are also equal in length with minimum outliers indicating that the sample bootstraped data is approximately normally distributed.

[2 marks] Based on Part a) and b), construct a 90% confidence interval for the mean income for this company and interpret this interval.

Ans :

As the bootstraped sample means in part b provided us a with normally of the distribution, we can use a parametric based approach using t-test to calculate the 90% confidence interval.

```r
ttest <- t.test(bootstrap_samples$stat, conf.level = 0.90)
ttest
```

```

    One Sample t-test

data:  bootstrap_samples$stat
t = 2842.3, df = 3999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 105283.2 105405.1
sample estimates:
mean of x
 105344.1
```

We are 90% confident that the population mean income for is between $1,05,283.2 and $1,05,405.1.

d) [2 marks] What does this imply about using Shapiro-Wilks to assess the Central Limit Theorem's normality condition?

Ans :

```r
shapiro.test(bootstrap_samples$stat)
```

```

	Shapiro-Wilk normality test

data:  bootstrap_samples$stat
W = 0.99929, p-value = 0.1191
```

In part a , we used shapiro test on original data, and it showed that the original data was not normally distributed. Now we applied shapiro test on the bootstraped sample means of the data , and it clearly shows that the data of sample means from the bootstrap sample is normally distributed.

Hence we can conclude that shaprio test is used to access the normality of the original data, but according to the Central Limit Theorem (CLT), even if the original data is not normally distributed, the sample means will be approximately normal if the sample size is large enough. And to verify this, we do not need to perform shaprio test on the sample means, as CLT will ensure their normality.

Question 2- Parametric Hypothesis Test [9 marks]

A Harris Poll asked Americans whether states should be allowed to conduct random drug tests on elected officials. Of 21,355 respondents, 16,870 said "yes".

a) [1 mark] What is the population?

Ans : The population is : All American Residences

b) [1 mark] What is the parameter of interest?

Ans : proportion of American residences who responded yes to conduct random drug tests on elected officials.

c) [1 mark] Estimate the parameter in part b) with the data.

Ans :

```r
n_yes <- 16870
total <- 21355

p <- n_yes/total
p
```

```
[1] 0.7899789
```

The estimated parameter (sample proportion) is 79%

d) [6 marks] The press release for this poll stated that "less than 80% of Americans agree with random drug testing of elected officials". Is this supported by the data? Perform a parametric hypothesis test to answer this question using a 5% level of significance (= 0.05).

Ans :

Hypothesis Testing :

- Hypothesis :
  - Ho : p = 0.80 (True Proportions of Americans who says Yes is 80%)

  - Ha : $p < 0.80$ (True Proportion of Americans who says Yes is less than 80%)
- Assumptions :
  - Random Sampling : As it is a poll based survey, we can assume that it was randomly sampled.
  - Independence : As the sample size is large enough , and it is also less than 10% of the entire American Population.
  - Sample size and Normality : The sample size is large enough ($np > 10$) npo = 21355 x 0.80 = 17,084 > 10 (greater than 10), n(1-po) = 21,355 * 0.20 = 4,271 (greater than 10) As both the conditions are satisfied , we can say that the proportion is approximately Normally Distributed.
- Calculating the test statistic (in this case we can use Z-score) :

```
n <- 21355  # Total responses
x <- 16870  # "Yes" responses
p_hat <- x / n
p0 <- 0.80
alpha <- 0.05

# Z-Score
z <- (p_hat - p0) / sqrt((p0 * (1 - p0)) / n)
z
```

```
[1] -3.661036
```

- Finding the p-value for the calculated statistic :

```
p_value <- pnorm(z)
p_value
```

```
[1] 0.0001255988
```

- Accepting/Rejecting Null Hypothesis Ho based on p-value :

```
if (p_value < alpha) {
  result <- "Reject the null hypothesis"
} else {
  result <- "Accept null hypothesis"
}


cat("Sample proportion (p̂):", p_hat, "\n")
```

```
Sample proportion (p̂): 0.7899789
```

```r
cat("Test statistic (z):", z, "\n")
```

```
Test statistic (z): -3.661036
```

```r
cat("P-value:", p_value, "\n")
```

```
P-value: 0.0001255988
```

```r
cat("Decision:", result, "\n")
```

```
Decision: Reject the null hypothesis
```

As the p-value is much smaller than the significance level , hence we reject the null hypotheis.

- Conclusion : There is enough evidence to support that less than 80% of Americans agree for random drug test.

Question 3: Non-Parametric Inference [8 marks]

A sociologist is studying the impact of private health insurance on well-being for Canadians, and gathers a small survey of = 250 Canadian's with some limited data on their health. The data is contained in a spreadsheet named r3data.csv (on Blackboard). The data description is below:

- **Name:** first name of survey participant
- **Age:** discrete numerical variable, age of participant, in years
- **Insurance:** categorical variable, with "Private Insurance" meaning the participant has private health insurance, otherwise, "No Private Insurance"
- **Income_Bracket:** categorical variable, average annual gross income decile, in tens of thousands of dollars
- **Prescriptions:** categorical variable, with "TRUE" meaning the participant has current, regularly filled, prescription medications prescribed by a medical professional, that insurance may cover

a) [1 mark] What is the sociologist's population of interest?

Ans : The population of interest is all Canadian residents.

b) [1 mark] What is the sociologist's parameter of interest?

Ans : The parameter of interest is true proportion of Canadians who have private health insurance.

c) [3 marks] The sociologist has been told that at least 70% of Canadians have private health insurance. Does her data support this claim? Perform a parametric bootstrap (infer-based) hypothesis test to check this claim.

Ans :

Loading the data

```r
library(readr)
data <- read_csv("r3data.csv")
```

```
Rows: 250 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr (2): Name, Insurance
dbl (2): Age, Income_Bracket
lgl (1): Prescriptions

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
str(data)
```

```
spc_tbl_ [250 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Name          : chr [1:250] "Antonio" "Jakiyah" "Courtney" "Bailey" ...
 $ Age           : num [1:250] 55 51 35 53 62 50 51 39 37 58 ...
 $ Insurance     : chr [1:250] "Private Insurance" "Private Insurance" "Private Insurance" "P
 $ Income_Bracket: num [1:250] 100 140 80 40 60 100 140 60 140 60 ...
 $ Prescriptions : logi [1:250] TRUE FALSE TRUE TRUE TRUE FALSE ...
 - attr(*, "spec")=
  .. cols(
  ..   Name = col_character(),
  ..   Age = col_double(),
  ..   Insurance = col_character(),
  ..   Income_Bracket = col_double(),
  ..   Prescriptions = col_logical()
  .. )
 - attr(*, "problems")=<externalptr>
```

```r
head(data,5)
```

```
# A tibble: 5 x 5
  Name       Age Insurance         Income_Bracket Prescriptions
  <chr>    <dbl> <chr>                      <dbl> <lgl>
1 Antonio     55 Private Insurance            100 TRUE
2 Jakiyah     51 Private Insurance            140 FALSE
3 Courtney    35 Private Insurance             80 TRUE
4 Bailey      53 Private Insurance             40 TRUE
5 Micheal     62 Private Insurance             60 TRUE
```

Calculating the proportion of population who has Private Insurance

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v stringr   1.5.1
v forcats   1.0.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
observed_proportion <- data %>%
  filter(Insurance == "Private Insurance") %>%
  summarise(prop = n() / nrow(data)) %>%
  pull(prop)
```

Bootstrap hypothesis test (p = 0.7)

```r
bootstrap_test <- data %>%
  specify(response = Insurance, success = "Private Insurance") %>%
  hypothesize(null = "point", p = 0.7) %>%
  generate(reps = 1000, type = "draw") %>%
  calculate(stat = "prop")
```

Calculation of p-value

```r
p_value <- bootstrap_test %>%
  summarise(p_value = mean(stat <= observed_proportion)) %>%
  pull(p_value)

cat("Observed Proportion:", observed_proportion, "\n")
```

```
Observed Proportion: 0.712
```

```r
cat("P-Value:", p_value, "\n")
```

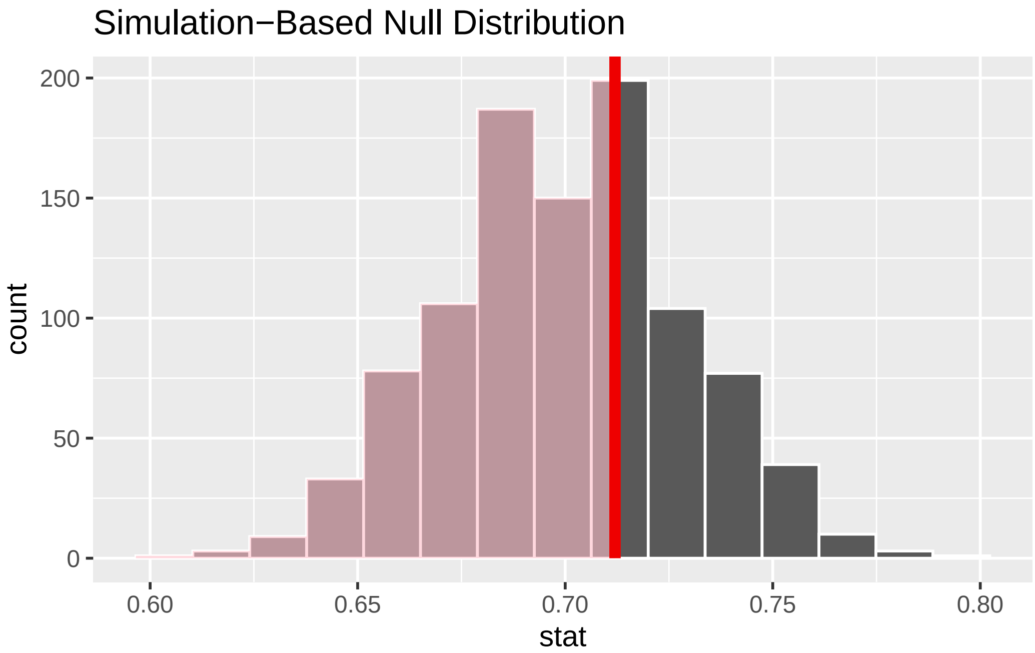
```
P-Value: 0.68
```

P value is : 0.682

d) [1 mark] Visualize the p-value for this test using the built-in infer visualization tool.

Ans :

```r
bootstrap_test %>%
  visualize() +
  shade_p_value(obs_stat = observed_proportion, direction = "left")
```



The shaded region in red represents the p-value.

e) [2 marks] Also using infer, compute a 91% confidence interval for the true underlying population proportion of Canadians who have private health insurance.

Ans :

Confidence Interval (91%)

```
confidence_interval <- data %>%
  specify(response = Insurance, success = "Private Insurance") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_confidence_interval(level = 0.91, type = "percentile")


confidence_interval
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1    0.668     0.76
```

Conclusion

We are 91% confident that the true proportion of Canadians having private health insurance lies between 66.4% - 76%.