STAT 630: Homework 5

Ly Nguyen

Due: October 26th

Note: You may use any R functions to compute p-values and confidence intervals.

1) A researcher conducted a study to investigate whether Nissan charges more for their sedan models than Toyota. Assume the prices of sedans at both companies are approximately normal. The researcher randomly selected 8 sedans from Nissan and 8 sedans from Toyota. Answer the following questions to determine if data provide convincing evidence that, **on average**, Nissan charges a larger amount than Toyota for sedans?

Summary statistics of the purchase prices and the differences (Nissan - Toyota) are shown in the table below.

	n	Mean	Standard Deviation
Nissan	8	$25,656.88	$9,726.72
Toyota	8	$25,058.38	$9,659.57
Difference	8	$598.50	$482.13

Perform a 5-step hypothesis to answer the question of interest. Show all steps below.

1) Write the hypotheses.

H_0 : Nissan charges an equal amount compared to Toyota does for sedans, on average.

H_A : Nissan charges a larger amount than Toyota for sedans, on average.

2) Check conditions.

- Independence:

The selected 8 sedans from Nissan and 8 sedans from Toyota are randomly selected.

Nissan and Toyota are 2 separated companies, so their charges for their products are independent.

- Normality:

Assume the prices of sedans at both companies are approximately normal.

n1: number of cars selected from Nissan

n2: number of cars selected from Toyota.

```
n1 <- 8
n2 <- 8
```

$n_1 = 8$

$n_2 = 8$

n1, n2 < 30 and data are approximately normal.

3) Calculate test statistic.

```r
xbar1 <- 25656.88
xbar2 <- 25058.38


s1 <- 9726.72
s2 <- 9659.57

stat <- xbar1 - xbar2
stat
```

```
## [1] 598.5
```

```r
null_value <- 0
se <- sqrt(s1^2/n1 + s2^2/n2)
se
```

```
## [1] 4846.602
```

```r
t_stat <- (stat - null_value) / se
t_stat
```

```
## [1] 0.1234886
```

4) Calculate p-value.

```r
df <- min(c(n1, n2)) - 1
df
```

```
## [1] 7
```

```r
p_val <- round(pt(t_stat, df = df, lower.tail = FALSE),4)
p_val
```

```
## [1] 0.4526
```

$p - value >$ the significance level 0.05

5) Make a decision and conclude in the context of the problem.

Decision: Fail to reject H_0
Conclusion: We do not have enough evidence that Nissan charges a larger amount than Toyota for sedans, on average.

99% Confidence Interval

```r
stat + c(-1, 1) * qt(0.995, df = df) * se
```

```
## [1] -16362.1  17559.1
```

We are 99% confidence that the true difference in average amount that Nissan charges for sedan minus the amount Toyota does is between -16,362.1 and 17,559.1

2) List two values that affect with width of a confidence interval *for a population proportion* specifically. Explain how increasing or decreasing these values changes the width of a confidence interval.

There are two values that affect with width of confidence interval: sample size and confidence level.

*Sample size: if the sample size increases, the confidence interval decreases. Explanation: Since if sample sizes provide more data, it means that the sample provide more precise estimates of the population proportion. So, the estimate would be closer to the true population proportion, as a result, margin of error decreases, leading to a narrower interval. Conversely, sample size decreases, confidence interval increases.

- Confidence Level: if the confidence level increases, the width of the confidence interval increases. Explanation: Since a higher confidence level requires a larger range of values that could potentially contain the true population proportion. The interval, therefore, needs to be larger to accommodate this increased level of certainty.

3) The following dataset contains auction data from Ebay for the game Mario Kart for the Nintendo Wii. This data was collected in early October 2009 and is available in the "openintro" package.

Using a confidence interval (instead of a p-value), determine if there is a significant difference in **the price of new versus used** Mario Kart games for Nintendo Wii's. Make sure to 1) write the hypothesis, 2) check conditions, 3) compute the confidence interval, 4) explain how you used the confidence interval to make a decision, and 5) conclude in the context of the problem.

```
library(openintro)
data("mariokart")
data("cle_sac")
library(dplyr)
library(openintro)
```

Step 1) write the hypothesis

H_O : There is no significant difference in the price of new versus used Mario Kart games for Nintendo Wii.

H_A : There is a significant difference in the price of new versus used Mario Kart games for Nintendo Wii.

Step 2) Check conditions

- Independence:

Assuming the price of new and used Mario Kart games for Nintendo Wii is randomly selected.

- Normality:

$n_1 > 30$

$n_2 > 30$

```
price_new <- mariokart %>%
  filter(cond == "new", !is.na(total_pr)) %>%
  select(total_pr) %>%
  pull()
price_new
```

```
##  [1] 51.55 45.50 44.00 71.00 45.00 53.99 54.99 56.01 48.00 56.00 46.71 46.00
## [13] 55.99 53.98 64.95 50.50 55.00 47.00 63.99 53.76 46.03 51.99 55.99 53.99
## [25] 59.88 51.99 53.99 58.00 53.99 47.00 41.50 56.00 64.95 54.99 64.00 54.70
## [37] 49.91 40.10 52.59 48.99 66.44 63.50 47.00 53.76 46.00 57.50 75.00 45.99
## [49] 45.00 49.75 56.00 46.00 61.00 62.89 46.00 64.95 58.98 47.70 54.51
```

```
price_used <- mariokart %>%
  filter(cond == "used", !is.na(total_pr)) %>%
  select(total_pr) %>%
  pull()
price_used
```

```
##  [1]  37.04  37.02  47.00  50.00  43.33  46.00 326.51  31.00  46.50  34.50
## [11]  36.00  40.00  43.00  31.00  41.99  49.49  41.00  44.78  47.00  44.00
## [21]  42.25  46.00  41.99  39.00  38.06  46.00  28.98  36.00  43.95  32.00
## [31]  40.06  48.00  36.00  31.00  30.00  38.10 118.50  61.76  40.00  64.50
## [41]  49.01  40.10  49.00  48.00  38.00  45.00  41.95  43.36  45.21  65.02
## [51]  45.75  36.00  47.00  43.00  35.99  54.49  46.00  31.06  55.60  44.00
```

```
## [61]  38.26  51.00  42.00  55.00  33.01  43.00  42.55  52.50  48.92  40.05
## [71]  50.00  47.00  41.00  34.99  49.00  36.99  44.00  41.35  37.00  39.00
## [81]  40.70  39.51  52.00  38.76
```

```
n1 <- length(price_new)
n1
```

```
## [1] 59
```

```
n2 <- length(price_used)
n2
```

```
## [1] 84
```

Step 3) compute the confidence interval

```
# Calculate the sample means
mean_new <- mean(price_new)
mean_used <- mean(price_used)

# Calculate the standard errors
se_new <- sd(price_new) / sqrt(length(price_new))
se_used <- sd(price_used) / sqrt(length(price_used))

# Calculate the degrees of freedom
df1 <- min(length(price_new) - 1, length(price_used) - 1)
df1
```

```
## [1] 58
```

```
# Calculate test statistic
stat1 <- mean_new - mean_used

se1 <- sqrt(se_new^2/n1 + se_used^2/n2)
se1
```

```
## [1] 0.4096519
```

```
# 99% Confidence Interval

CI <- stat1 + c(-1, 1) * qt(0.995, df1) * se1
CI
```

```
## [1] 5.531562 7.713603
```

Step 4) explain how you used the confidence interval to make a decision

Since the confidence interval does not include zero, in this case it implies that the difference in means is statistically significant at the chosen confidence level (typically 99% confidence level is used). **So, we reject H_0.**.

Step 5) conclude in the context of the problem.

We have enough evidence that there is a significant difference in the price of new versus used Mario Kart games for Nintendo Wii.

4) In a recent poll of 500 13-year-olds, many indicated to enjoy their relationships with their parents. Suppose that 200 of the 13-year olds were boys and 300 of them were girls. We wish to estimate the difference in proportions of 13-year old boys and girls who say that their parents are very involved in their lives. In the sample, 93 boys and 172 girls said that their parents are very involved in their lives.

a. Calculate a 95% confidence interval for the difference in proportions (proportion of boys minus proportion of girls)? Include any R code used.

n_boy: 200 boys

n_girl: 300 girls

p_girl: proportion of the girls who say that their parents are very involved in their lives.

p_boy: proportion of the boys who say that their parents are very involved in their lives.

```
n_boy <- 200
n_girl <- 300
p_boy <- 93 / n_boy
p_girl <- 172 / n_girl

# Calculate the standard error
se2 <- sqrt(p_boy*(1-p_boy)/n_boy + p_girl*(1-p_girl)/n_girl)

# Calculate the Z-score for the 95% confidence level
z <- qnorm(1 - (1 - 0.95) / 2)

# Calculate the margin of error
margin_of_error <- z * se2

# Calculate the confidence interval
lower_bound <- round(((p_boy - p_girl) - margin_of_error), 4)
lower_bound
```

```
## [1] -0.1973
```

```
upper_bound <- round(((p_boy - p_girl) + margin_of_error), 4)
upper_bound
```

```
## [1] -0.0194
```

b. Interpret your interval calculated above in the context of the problem.

The calculated 95% confidence interval for the difference in proportions between boys and girls who say their parents are very involved in their lives is:

-0.1973 $\leq$ Proportion of Boys - Proportion of Girls $\leq$ -0.0194

Interpreting this in the context of the problem:

The negative confidence interval reveals that the girls are more likely than the boys to report that their parents are very involved in their lives. Besides, the true difference between the proportions of boys and girls who report high parental involvement is at least -0.1973 and highest at -0.0194 in favor of the girls.