Activity 6 - Salvin Chowdhury

Due Date: April 30, 2025

Problem 1:

A school counselor suspects that students' test scores may be affected by the site at which they take the test. An independent random sample of 31 test scores are collected from Test Center A and 40 from Test Center B.

You will be conducting a hypothesis test with significance level 0.05 to determine if there is a difference in mean scores between the two test centers.

The data is recorded in "Activity6ExamScores.csv" which can be found in Canvas. Be sure to run the code chunk below so that you can use the csv file within this markdown file.

```
# reading in the .csv file into a new data frame
ExamScores <- read.csv("C:/GitHub/StatisticsInR/wk_14/Activity6ExamScores.csv")
```

(a) First, determine the null and alternative hypotheses. **Answer for (a):** The Null Hypothesis is:

- $H_0 : \mu_A = \mu_B$

The Alternative Hypothesis is: - $H_1 : \mu_A - \mu_B \neq 0$

The Null Hypothesis is that there is no difference in mean scores between the two test centers The Alternative Hypothesis is that there is a difference in mean scores between the two test centers

(b) Take a look at the ExamScores data. You might notice that in its current state, performing any relevant calculations might not be so easy... To fix this, let's make two new datasets (one for each test center) using the subset() function - be sure to assign these as objects so you can continue to work with them. This function really just needs two inputs - the data you want to subset, and a logical expression indicating which rows to keep.

Note that when working with strings in R, be sure to put the string in between two quotation marks. Also note that the syntax for checking whether two objects are equal in R is a double equals sign "==". This step may require some "trial and error" - try a few ideas out and discuss with your neighbors first. Then, if you're still stuck of course I'll be happy to help you figure it out*

```
# creating a new data frame of just center A
center_A = subset(ExamScores, Center=="CenterA")
print(center_A)
```

```
##    Score  Center
## 2     84 CenterA
## 3     88 CenterA
## 7     87 CenterA
## 9     70 CenterA
## 11    72 CenterA
## 12    76 CenterA
## 13    83 CenterA
## 15    74 CenterA
## 19    79 CenterA
## 20    79 CenterA
## 26    85 CenterA
```

```
## 27    85 CenterA
## 29    97 CenterA
## 30    60 CenterA
## 31    64 CenterA
## 36    76 CenterA
## 38    88 CenterA
## 41    89 CenterA
## 43    94 CenterA
## 50    91 CenterA
## 53    73 CenterA
## 55    78 CenterA
## 58    86 CenterA
## 59    83 CenterA
## 60    74 CenterA
## 62    91 CenterA
## 64    82 CenterA
## 65    96 CenterA
## 66    95 CenterA
## 67    73 CenterA
## 68    82 CenterA
```

```r
# creating a new data frame of just center B
center_B = subset(ExamScores, Center=="CenterB")
print(center_B)
```

```
##    Score  Center
## 1     57 CenterB
## 4     79 CenterB
## 5     81 CenterB
## 6     78 CenterB
## 8     89 CenterB
## 10    82 CenterB
## 14    91 CenterB
## 16    63 CenterB
## 17    78 CenterB
## 18    79 CenterB
## 21    93 CenterB
## 22    93 CenterB
## 23    74 CenterB
## 24    87 CenterB
## 25    74 CenterB
## 28    66 CenterB
## 32    91 CenterB
## 33    62 CenterB
## 34    70 CenterB
## 35    85 CenterB
## 37    72 CenterB
## 39    64 CenterB
## 40    80 CenterB
## 42    84 CenterB
## 44    64 CenterB
## 45    75 CenterB
## 46    75 CenterB
## 47    84 CenterB
## 48    84 CenterB
```

```
## 49    59 CenterB
## 51    82 CenterB
## 52    78 CenterB
## 54    66 CenterB
## 56    83 CenterB
## 57    85 CenterB
## 61    99 CenterB
## 63    70 CenterB
## 69    83 CenterB
## 70    85 CenterB
## 71    76 CenterB
```

(c) Now you're ready to perform the hypothesis test. You can do this in more of "manual" way (calculating the test statistic and degrees of freedom, and then using that to find the P-value), or by using the t.test() function to do this all at once - the choice is up to you.

```
# performing a t-test and printing out the results
result_one <- t.test(center_A$Score, center_B$Score)
print(result_one)
```

```
##
##  Welch Two Sample t-test
##
## data:  center_A$Score and center_B$Score
## t = 1.6317, df = 67.16, p-value = 0.1074
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8353153  8.3191862
## sample estimates:
## mean of x mean of y
##  81.74194  78.00000
```

(d) Based on your result from part (c), is the null hypothesis rejected? State a conclusion/interpretation of your results in the context of this problem.

Answer for (d): The p-value we get here is 0.1074. Since our signifigance level is 0.05, we fail to reject the null hypothesis. This means that there isn't enough evidence to support the claim that there is a difference in mean scores between the two test centers.

Problem 2

The Deloitte employment survey asked a sample of HR executives how their company planned to change its workforce over the next 12 months. A categorical response variable showed the following three options: Add employees, No change, and Lay off employees.

Another categorical variable indicated if the company was private or public. A researcher is interested in finding out whether the percentage of private companies that plan to add employees is higher than the percentage of public companies that wish to do so.

We will investigate this using a hypothesis test with significance level 0.01. Sample data for 180 companies are recorded in "Activity6WorkforcePlan.csv" which can be found in Canvas. Be sure to run the code chunk below so that you can use the csv file within this markdown file.

```
# creating a new data frame for the work force
WorkforcePlan <- read.csv("Activity6WorkforcePlan.csv")
```

(a) State the null and alternative hypotheses. **Answer for (a):**
The Null Hypothesis is:

- $H_0 : p_1 - p_2 = 0$

The Alternative Hypothesis is:
- $H_1 : p_1 - p_2 > 0$

The Null Hypothesis is that there is no difference in percentage of private companies than plan to add employees and percentage of public companies The Alternative Hypothesis is that the percentage of private companies that plan to add employees is higher than the percentage of public companies

(b) What are each of the sample sizes and proportions in question (that is, m, n, and the two "phats"). Hint: using the table() function on each of the columns of your dataset may help you here (although this is certainly not the only method to help obtain this information).

```
# the number of private companies
number_of_private <- sum(!is.na(WorkforcePlan$Private) & WorkforcePlan$Private != "")
print(number_of_private)
```

```
## [1] 72
```

```
# the number of public companies
number_of_public <- sum(!is.na(WorkforcePlan$Public) & WorkforcePlan$Public != "")
print(number_of_public)
```

```
## [1] 108
```

```
# number of private companies that will add employees
public_add <- sum(WorkforcePlan$Public == "Add", na.rm=TRUE)
print(public_add)
```

```
## [1] 32
```

```
private_add <- sum(WorkforcePlan$Private == "Add", na.rm=TRUE)
print(private_add)
```

```
## [1] 37
```

```
# finding the proportions
p_hat_1 <- public_add / number_of_public
p_hat_2 <- private_add / number_of_private

# printing out the results
print(p_hat_1)
```

```
## [1] 0.2962963
```

```
print(p_hat_2)
```

```
## [1] 0.5138889
```

Answer for (b): - m = 108 - n = 108 - p_hat_1 = 0.2962963 - p_hat_2 = 0.5138889

(c) Use the prop.test() function to find the test statistic and P-value. You will need the following arguments: x, n, alternative, conf.level, correct. x is a vector of the number of successes from each sample. n is a vector of the number of trials in each sample. alternative and conf.level are just like in the t.test function. Use correct = FALSE as the final argument.

```
# performing the prop.test()
new_result <- prop.test(
    c(private_add, public_add),
    c(number_of_private, number_of_public),
```

```
    alternative="greater",
    correct=FALSE
)

# printing out the result
print(new_result)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(private_add, public_add) out of c(number_of_private, number_of_public)
## X-squared = 8.6526, df = 1, p-value = 0.001633
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.09671943 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5138889 0.2962963
```

(d) Based on your result from part (c), is the null hypothesis rejected? State a conclusion/interpretation of your results in the context of this problem. **Answer for (d):** Given that the p-value is 0.001633, and that our signifigance level is 0.01, this means that we reject the null hypothesis. As a result, this means that there is enough statistical evidence to prove that that the percentage of private companies that plan to add employees is higher than the percentage of public companies.

Problem 3

Thirty cars were equipped with radial tires and driven over a test course. Then the same 30 cars (with the same drivers) were equipped with regular belted tires and driven over the same course. After each run, the cars' gas economy (in km/l) was measured. We wish to determine if there is evidence that radial tires produce better fuel economy than regular belted tires. We will do so by performing a paired t-test on the difference of means.

The relevant data is contained in "Activity6Tires.csv" which can be found in Canvas. Be sure to run the code chunk below so that you can use the csv file within this markdown file.

```
# reading in the tires data frame
Tires <- read.csv("Activity6Tires.csv")
```

(a) Consider the "D_i" values as radial fuel economy - belted fuel economy. State the null and alternative hypotheses. **Answer for (a):**

Null Hypothesis: - $H_0 : \mu_D = 0$

Alternative Hypothesis: - $H_1 : \mu_D > 0$

The Null Hypothesis is that there is no difference in the fuel economy between radial and regular belted tires The Alternative Hypothesis is that the radial tires have a better fuel economy than the regular belted tires

(b) Use the t.test function to find the test statistic and P-value (be sure to specify the appropriate arguments in the function).

```
# performing a pairwise t-test
result_two <- t.test(Tires$Radial, Tires$Belted, paired=TRUE, alternative='greater')
print(result_two)
```

```
##
##  Paired t-test
```

```
##
## data:  Tires$Radial and Tires$Belted
## t = 3.0293, df = 29, p-value = 0.002555
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.04683793        Inf
## sample estimates:
## mean difference
##       0.1066667
```

(c) Based on your result from part (b), is the null hypothesis rejected? State a conclusion/interpretation of your results in the context of this problem. **Answer for part (c):**

As the p-value we get is 0.002555, and comparing it to the signifigance level of 0.05, we reject the null hypothesis as there is statistical evidence that the radial tires have a better fuel economy than the regular belted tires