Practice Session 7

Intro

The focus of this practice session will be to perform hypothesis tests for the difference of two or more means. We will look at various test statistics that can be used. We will also conduct hypothesis testing for correlations.

Question 1: Exercise

A study is interested in whether the mean exercise hours differs between male and female students. Conduct a permutation test to see if there is a difference. Use the data set `ExerciseHours` from the `Lock5Data` library. Make sure to extract the appropriate variables from the data frame.

1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

2.) Create a boxplot to describe hours of exercise for `female` versus `male`.

```
# your code here
```

3.) Find some favorites statistics of `Exercise hours` for `female` and `male` students. You might find the function: `mosaic::favstats` useful. *Hint*: you can search online for the function's arguments.

```
#your code here
```

4.) Subset the data `ExerciseHours` to two groups: `F` and `M`.

```
#your code here
```

5.) Compute the observed statistic (mean difference of exercise hours for Female and Male).

```
#your code here
```

6.) Create null hypothesis distribution

- a.) Shuffle the two groups of `female` and `Male` into two samples, and find the mean difference of the two shuffled samples.
- b.) Create the Null hypothesis Distribution using `do_it()` function.
- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed mean difference` using the `abline()` function.

```
# your code here
```

7.) Calculate p-value

```
# your code here
```

8.) Make a decision and state your conclusion:

```
#your code here
```

Question 2: Caffeine Taps

A sample of male college students were asked to tap their fingers at a rapid rate. The sample was then divided at random into two groups of ten students each. Each student drank the equivalent of about two cups of coffee, which included about 200 mg of caffeine for the students in one group but was decaffeinated coffee for the second group. After a two hour period, each student was tested to measure finger tapping rate (taps per minute). The goal of the experiment was to determine whether caffeine produces an increase in the average tap rate.

1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

2.) Create a boxplot to describe hours of exercise for `Caffeine` versus `No Caffeine`.

```
# your code here
```

3.) Find some favorite statistics to visualize the number of taps for the `Caffeine` and `No Caffeine` group.

```
#your code here
```

4.) Subset the data `CaffeineTaps` to two groups: `Caffeine` and `NoCaffeine`.

```
#your code here
```

5.) Compute the observed statistic (mean difference of tap number for the two groups).

```
#your code here
```

6.) Create null hypothesis distribution

- a.) Shuffle the two groups of `Caffeine` and `No Caffeine` into two samples, and find the mean difference of the two shuffled samples.
- b.) Create the Null hypothesis Distribution using `do_it()` function.
- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed mean difference` using the `abline()` function.

```
# your code here
```

7.) Calculate p-value

```
# your code here
```

8.) Make a decision and state your conclusion:

```
#your code here
```

Question 3: Hypothesis Testing for Correlations

Do more home runs mean more wins? In other words, is there a positive correlation between the number of home runs a team hits and the number of wins? Test this hypothesis using a permutation test. Data from the 2019 Major League Baseball (MLB) season is available in the `BaseballHits2019` dataset in the `Lock5Data` library. Make sure to extract the appropriate columns from the data.

1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

2.) Create a scatterplot to visualize the association between `HomeRuns` and `Wins`.

```
# your code here
```

3.) Compute the observed statistic (correlation between `HomeRuns` and `Wins`).

```
#your code here
```

4.) Create null hypothesis distribution

- a.) Shuffle the variables `HomeRuns` and `Wins` into two new variables, and find the correlation between these two new shuffled variables.
- b.) Create the Null hypothesis Distribution using `do_it()` function.
- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed correlation` using the `abline()` function.

```
# your code here
```

5.) Calculate p-value

```
# your code here
```

6.) Make a decision and state your conclusion:

```
#your code here
```

Question 4: Comparing Multiple Samples: The Mean Absolute Deviation (MAD) Statistic

The mean absolute deviation (MAD) statistic is a statistic that we can calculate to compare differences among more than two groups. Suppose we had 4 groups. Then the MAD statistic is given by:

$\text{MAD} = \frac{|\bar{x}_1-\bar{x}_2|+|\bar{x}_1-\bar{x}_3|+|\bar{x}_1-\bar{x}_4|+|\bar{x}_2-\bar{x}_3|+|\bar{x}_2-\bar{x}_4|+|\bar{x}_3-\bar{x}_4|}{6}$

The Lock5Data library contains the data set `TextbookCosts`. Conduct a permutation test to see if the mean textbook cost differs among subjects. *Hint:* Follow the steps that you did for the previous exercises, except now use the `get_MAD_stat()` function to compute your observed statistic and generate your null distribution.

Question 5: Non-parameteric test using vacccine antibodies (Kruskal-Wills test)

We will use data on `Antibodies` (in g/ml) production after receiving a `Vaccine` (Vaccine A, Vaccine B, Vaccine C). A hospital administered three different vaccines to 6 individuals each and measured the antibody presence in their blood after a chosen time period. The data is saved in `patient_vaccine.csv`.

We walk you through testing for the difference between the three groups of vaccines using a different method than in class, it is called the Kruskal-Wills test.

1) Create a boxplot to show the three vaccines variation in terms of the antibodies.

```
#your code here
```

2.) Write in words the `null hypothesis` and the `alternative hypothesis`.

```
#your code here
```

3.) Let prepare your data. Rank your data from all groups together in one column, name it `ranks`. *hint*: you can use function `rank`.

```
#your code here
```

4.) Sum the ranks for each group of the `Vaccine`. Reports those sums results.

```
#your code here
```

5.) Calculate the test statistic, `H` of the Kruskal-Wills test given by the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Where :

- N is the total sample size
- k is the number of groups we are comparing.
- $ {n_i} $ is the sum of ranks for group i.
- $ {R_i} $ is the sample size of group i.

```
#your code here
```

6.) Compare the test statistics`H` to the critical cutoff determined by the critical value `chi-square`. *hint*: from the chi-square table, find the chi-square `critical value` with degrees of freedom `df= k-1` .

```
#your code here
```

7.) Make Judgement about your hypothesis within the context.

```
#your code here
```

Question 6: Compare Mad method to Kruskal-Wills test

Repeat Question 5 with the randomization method `MAD` and compare your results. What is your reflection.

```
#your code here
```