Practice Session 10

Part 1

Chi-Square Test for Categorical Variables

$\chi^2 = \sum \frac{(\text{ Observed } - \text{ Expected })^2}{\text{Expected}}$

Question 1 : Chi-Square (χ^2) test - What Type of Ice Cream?

Sixty people were asked whether they preferred vanilla, chocolate, or strawberry ice cream and the results are shown in the table below. Perform a chi-square goodness-of-fit test to determine whether the flavors are equally popular.

Flavor	Frequency
Vanilla	28
Chocolate	23
Strawberry	9
TOTAL	60

1.) Write the hypothesis of this Chi-Square test. Do the hypothesized proportions have the same values?

2.) Find the df (degrees of freedom).

3.) Calculate the Chi-Square test statistic using the formula.

a) Find the expected values for each category.

b) Find the contribution of each category to the Chi-Square test statistic.

c) Calculate now the Chi-Square test statistic.

```
# your code here
```

4.) Find the p-value using `pchisq()` function. *hint*: find the function in help tab.

```
# your code here
```

5.) Make judgment about this test.

Question 2: Chi-Square (χ^2) test- Genetic Variation

Studies in genetics often involve chi-square tests. For one gene, we expect 25% of people to have the variant AA, 25% to have the variant BB , and 50% to have the variant AB . Observed counts of the three variants in one sample are shown. Do these counts provide evidence that the stated proportions are not right?

Variant	Frequency
AA	142
BB	121
AB	307
Total	570

1.) Write the hypothesis of this Chi-Square test.

2.) Find the df (degrees of freedom)

3.) Calculate the Chi-Square test statistic using the formula.

a) Find the expected values for each category.

b) Find the contribution of each category to the Chi-Square test statistic.

c) Calculate now the Chi-Square test statistic.

```
# your code here
```

4.) Find the p-value using `pchisq()` function. *hint*: find the function in help tab.

```
# your code here
```

5.) Make judgment about this test.

Part 2

ANOVA to compare means

Question 3: ANOVA- Hanging out with friends

A survey given to a sample of high school seniors in Pennsylvania in the data `PASeniors`. Two of the variables in the survey are `HangHours`, the number of hours per week spent hanging out with friends, and `SchoolPressure`, the amount of pressure felt due to schoolwork (None, Very little, Some, or A lot).

We wish to test whether the amount of school pressure felt by students is related to the mean time hanging out with friends.

1.) Prepare the data and clean it using `na.omit(my_dataframe)`.

```
library(SDS1000)
library(Lock5Data)
data(PASeniors)

cPASeniors<- na.omit(PASeniors)
```

2.) What is the explanatory variable? Is it categorical or quantitative? What is the response variable? Is it categorical or quantitative?

3.) Find the summary statistics for the groups of `SchoolPressure`, using function `mosaic::favstats( Response ~ grouping, data= )`. Which group has the largest mean and sd of the `HangHours` ?

```
#your code here
```

4.) What is the hypothesis?

5.) Use the function `aov(Response ~ grouping, data= )` to find the summary statistics for the ANOVA test. Indicate the F-stat and the p-value.

```
#your code here
```

6.) State the conclusion of the test in context.

Question 4: Running a randomization hypothesis test using an F-statistic for ANOVA- Hanging out with friends

Using the previous data `PASeniors`, and the two variables `HangHours` and `SchoolPressure`. Let test the previous hypothesis using the randomization distribution (test whether the amount of school pressure felt by students is related to the mean time hanging out with friends.

1.) Find the F-test-statistic using the function `get_F_stat()`. Compare it to question 3.

```
# your code here
```

2.) Create the null distribution and use `abline` to designate the F-stat.

```
# your code here
```

3.) Now calculate the p-values from the randomization and compare it to the previous p-value in question 3.

```
# your code here
```

4.) How many `k` groups are in the `PASeniors`? What is the sample size N? Create the F distribution density curve using the function `df()`, and the **two degrees of freedom**: `df 1= k-1` and the `df 2= n-k`.

Plot this density curve on the same histogram of the randomization distribution.

```
# your code here
```

Part 3

Inference for Linear Regression

Hypothesis testing can be done for more than sample means and proportions. We can also test hypotheses relating to regression parameters, like β_1, the slope, and β_0, the y-intercept. The hypotheses typically test whether the parameter is greater than zero, less than zero, or not equal to zero.

Question 5: Hypothesis Testing for Regression Slope β_1

Do triples lead to more wins? A "triple" in baseman occurs when a batter hits a ball in play, and is able to advance three bases (i.e., make it to third base). However, triples have become less common due to improvements in defensive positioning and outfielder speed. Perform a hypothesis test to see if hitting more triples is linearly associated with earning more wins. The data is available in the `BaseballHits2014` data set from the `Lock5Data` library.

a) Create a scattersplot to visualize the relationship between `Wins` and `Triples`

```
# your code here
```

b) State the null and alternative hypotheses using symbols

c) Fit the regression model with `Wins` as the outcome, and `Triples` as the predictor. Extract the slope coefficient by using the `coef()` function, and selecting the second value (e.g., `coef(my_model)[2]`).

```
# your code here
```

d) Create a null distribution by using the `do_it` function. The approach you will want to take is to fit a regression model inside the `do_it` call (maybe call it `curr_model`), and you will use `Wins` as the outcome, and a shuffled `Triples` as the predictor. You can shuffle the `Triples` variable using the `shuffle()` function. Extract the slope coefficient after fitting each model.

```
# your code here
```

e) Plot a histogram of your null distribution, and add a red vertical line at the observed slope

```
# your code here
```

f) Calculate the p-value by seeing the proportion of null values that are more extreme than the one you observed.

```
# your code here
```

g) State your conclusion.