Practice Session 10 Answers

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.

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

5.) Make judgment about this test.

Answers:

1.) The null hypothesis is that the flavors are equally likely to be preferred, while the alternative hypothesis is that the flavors are not equally likely to be selected.

In symbols:

- $H_0 : \pi_v = \pi_c = \pi_s = 1/3$
- H_a: Some $\pi_i \neq 1/3$

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

There are three categories so we have $df = 2$. Using a chi-square distribution with $df = 2$.

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

a) Find the expected values for each category.

The expected counts:

$60 \times 1/3 = 20$ in each category.

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

Contribution to the chi-square statistic, $(O - E)^2/E$, are shown as new columns in the table below.

Flavor	Observed	Expected	Contribution
Vanilla	28	20	3.2
Chocolate	23	20	0.45
Strawberry	9	20	6.05
TOTAL	60	60	9.7

c) Calculate now the Chi-Square test statistic.

We can see that the chi-square statistic is 9.7

$\chi^2 - test - statisti = \sum \frac{(\text{ observed } - \text{ expected })^2}{\text{expected}}$

4.) Find the p-value using `pchisq()` function. *hint*: find the function in help tab. Calculate now the Chi-Square test statistic.

We can see that the p-value is 0.0078

```
pchisq( 9.7, df= 2, lower.tail= F)
```

```
[1] 0.007828378
```

```
#0.007828
```

5.) Make judgment about this test.

The data provides strong evidence that the flavors are not equally preferred. (Note that the strongest contributor to the chi-square statistic is the count for strawberry, which is below the expected count).

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.

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

5.) Make judgment about this test.

Answers

1.) The null hypothesis is that the proportions of:

- $\pi_{AA} = 0.25$
- $\pi_{BB} = 0.25$
- $\pi_{AB} = 0.50$

While the alternative hypothesis is at least one of the null proportions is not correct.

2.) There are three categories so we have $df = 2$. Using a chi-square distribution with $df = 2$.

3.)

a) The expected counts are :

- For AA: 570 $\times$25% = 142.5
- For BB: 570 $\times$25% = 142.5
- For AB : 570 $\times$50% = 285

b) Find the contribution of each Category to the Chi^2 test statistic using $(O - E)^2/E$

Variant	Frequency	Observed	Contribution
AA	142	142.5	0.00
BB	121	142.5	3.24
AB	307	285	1.70
Total	570		4.94

c) We see that the chi-square test statistic is 4.94

4.) The p -value is 0.0845 .

```
pchisq( 4.94, df= 2, lower.tail= F)
```

```
[1] 0.08458486
```

```
#0.08458486
```

5.) At a 5% level, we do not reject H_0. These data do not contradict the hypothesized proportions for these three gene variants.

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.

Report the F-stat and the p-value.

```
#your code here
```

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

Answers:

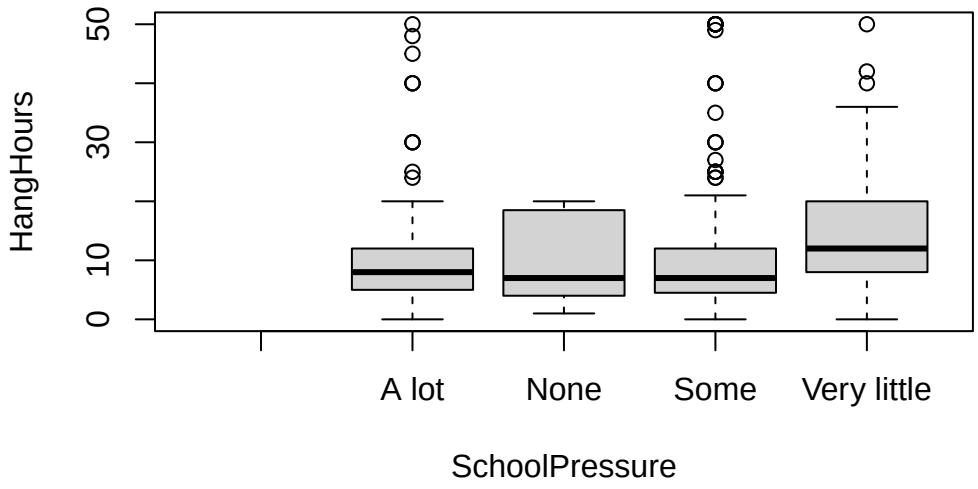
1.) prepare the data.

2.) The explanatory variable is amount of pressure felt from schoolwork, which is categorical. The response variable is number of hours per week hanging out with friends, which is 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` ?

The highest mean (15.91 hours per week hanging with friends) is for the group feeling 'Very little' pressure from schoolwork. The lowest mean (9.866 hours per week hanging out with friends) is for the group feeling 'A lot' of pressure from schoolwork.

```
#create boxplot for the data
boxplot(HangHours ~ SchoolPressure, data= cPASeniors)
```




```
# find the statistics for the groups
mosaic::favstats(HangHours ~ SchoolPressure, data= cPASeniors)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

	SchoolPressure	min	Q1	median	Q3	max	mean	sd	n	missing
1		NA	NA	NA	NA	NA	NaN	NA	0	0
2	A lot	0	5.000	8	12.00	50	10.34247	9.319012	146	0
3	None	1	4.000	7	17.75	20	10.08333	7.704288	12	0
4	Some	0	4.625	7	12.00	50	11.44481	12.136689	154	0
5	Very little	0	8.000	12	20.00	50	15.76744	12.096511	43	0

```
# create the ANOVA model
anova_model<- aov( HangHours ~ SchoolPressure, data= cPASeniors)


# summary statistics of the ANOVA model
summary(anova_model)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
SchoolPressure   3   1003   334.3   2.799   0.04 *
Residuals      351  41928   119.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.) H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a : Some $\mu_i \neq \mu_j$

5.) We see in the output that the F-statistic is 2.799 and the p-value is 0.04 .

6.) The p -value is small than 0.05, so we reject H_O. We have evidence of a difference in mean number of hours a week spent hanging out with friends depending on how much stress students feel from schoolwork.

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.

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

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

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.

Answers:

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

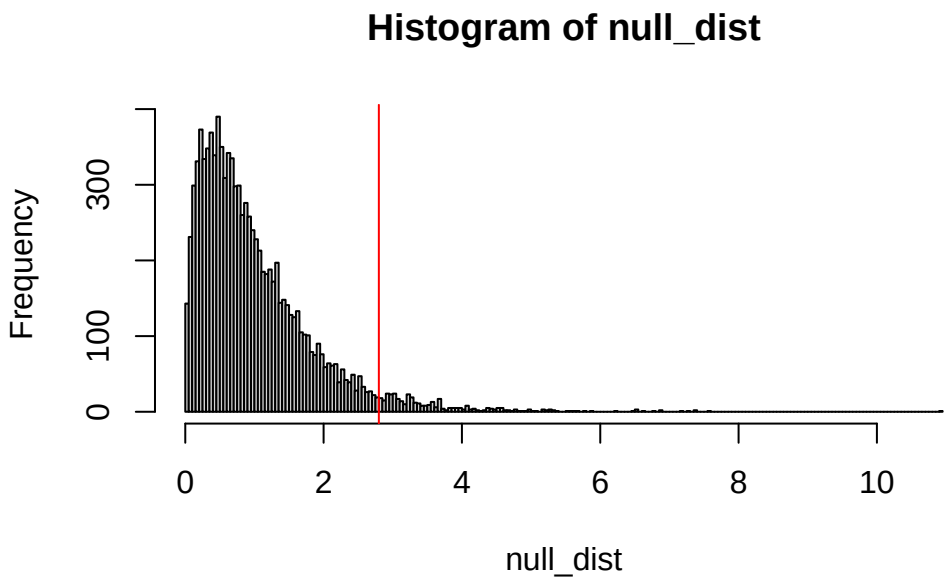
```
F_stat <- get_F_stat(cPASeniors$HangHours, cPASeniors$SchoolPressure)
F_stat
```

```
[1] 2.798637
```

The F-stat is 2.798 and it is the same as the previous F-stat found in Questions 3.

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

```
## create the null distribution

 null_dist <- do_it(10000) * {
shuffled_SchoolPressure <- shuffle(cPASeniors$SchoolPressure)
get_F_stat(cPASeniors$HangHours, shuffled_SchoolPressure)
}

null_hist<- hist(null_dist, breaks = 200)
# visualize the null distribution

hist(null_dist, breaks = 200)
abline(v = F_stat, col = "red")
```



3.) Now calculate the p-values from the randomization.

```
p_value <- pnull(F_stat, null_dist, lower.tail = F)
p_value
```

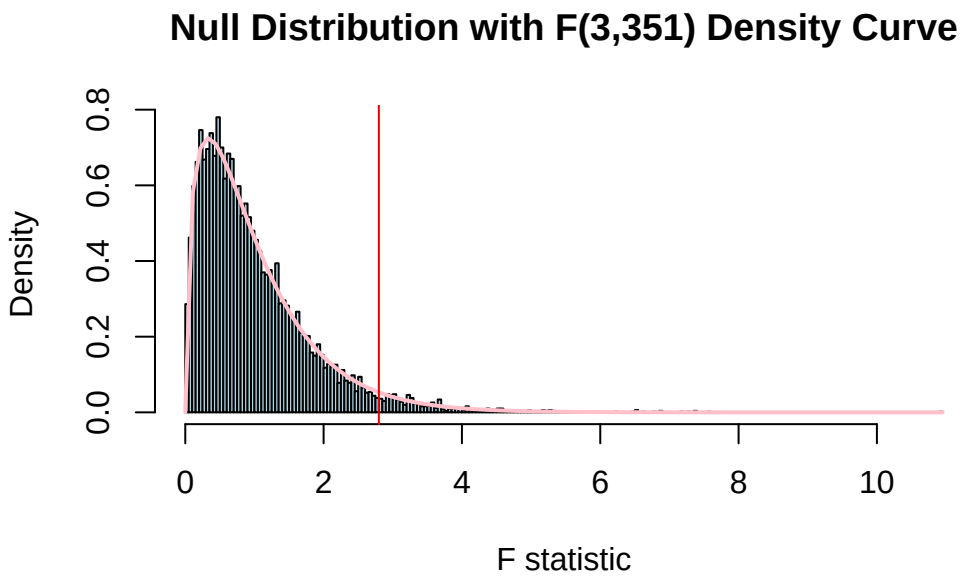
```
[1] 0.039
```

The p-value is 0.0439 which is the same as the p-value found in Question 3. (The answers may vary due to the randomization).

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** of the F-density are: `df 1= k-1` and the `df 2= n-k`.

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

```r
# Plot the F-distribution

k<- 4          # number of groups
n<- 355        # sample size

df1<- k-1      # degrees of freedom of the "Between Variation"
df2<- n-k      # degrees of freedom of the "Within Variation"


# Density histogram
hist(null_dist,
     breaks = 200,
     freq = FALSE,
     main = "Null Distribution with F(3,351) Density Curve",
     xlab = "F statistic",
     col = "lightblue")

# Overlay F-distribution density curve
curve(df(x, df1 = 3, df2 = 351),
      add = TRUE,
      col = "pink",
      lwd = 2)

abline(v = F_stat, col = "red2")
```



Part 3

Inference for Linear Regression

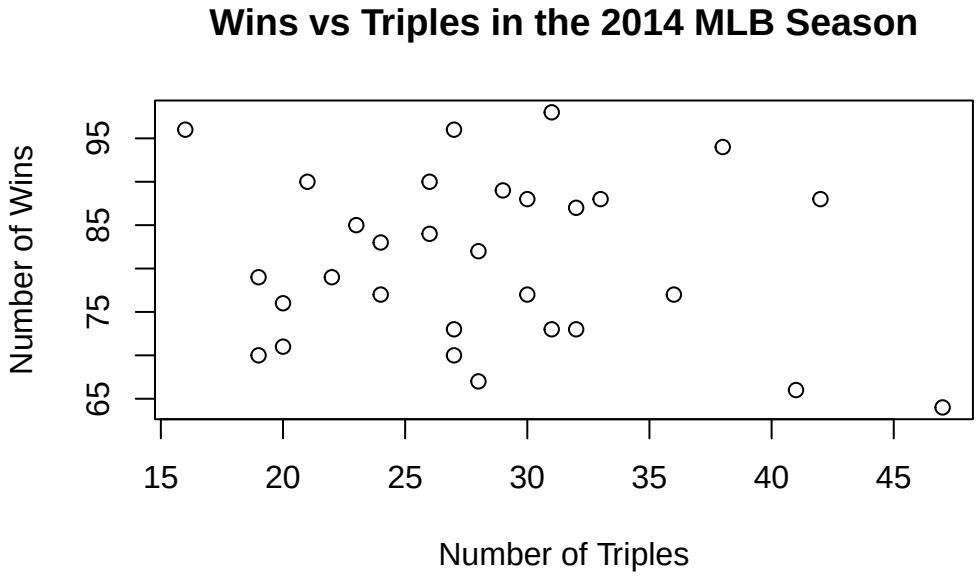
Hypothesis testing can be done for more than sample means and proportions. We can also test hypotheses relating to regression parameters, like β_1, the slope, and β_0, the y-intercept. The hypotheses typically test whether the parameter is greater than zero, less than zero, or not equal to zero.

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`

```
library(Lock5Data)
library(SDS1000)
plot(BaseballHits2014$Triples, BaseballHits2014$Wins, xlab = "Number of Triples",
     ylab = "Number of Wins", main = "Wins vs Triples in the 2014 MLB Season")
```



b) State the null and alternative hypotheses using symbols

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 > 0$

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]`).

```
triple_mod = lm(data = BaseballHits2014, Wins ~ Triples)
obs_beta1 = coef(triple_mod)[2]
obs_beta1
```

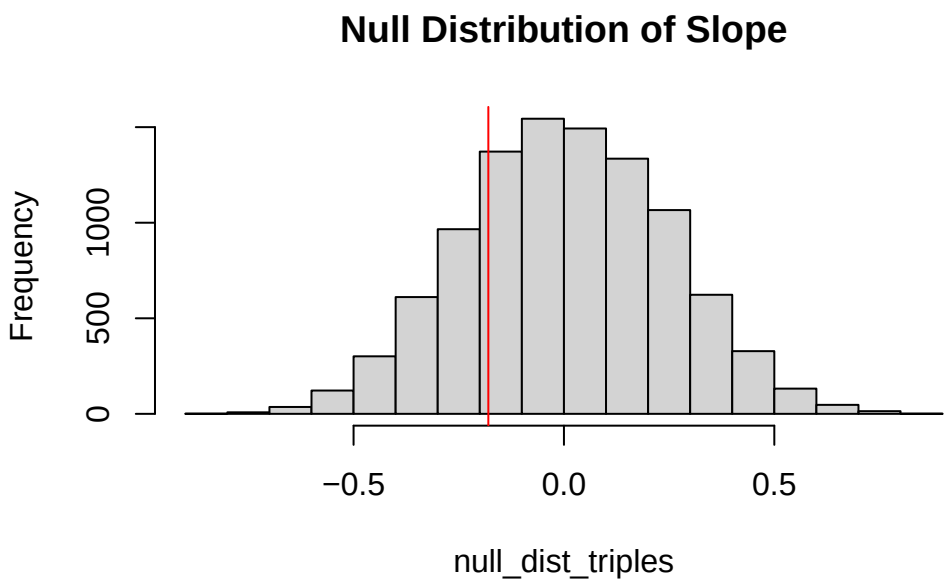
```
  Triples
-0.179683
```

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.

```
null_dist_triples = do_it(10000) * {
  curr_mod = lm(data = BaseballHits2014, Wins ~ shuffle(Triples))
  coef(curr_mod)[2]
}
```

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

```
hist(null_dist_triples, main = "Null Distribution of Slope")
abline(v = obs_beta1, col = "red")
```



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

```
pnull(obs_beta1, null_dist_triples, lower.tail = F)
```

```
[1] 0.7701
```

g) State your conclusion.

Since our p-value is greater than 0.05, we will fail to reject the null hypothesis. We therefore do not have evidence that hitting more triples is linearly associated with earning more wins.