Logo

Description automatically generated

ALY6015 Module 2 Project – R Practice

**Authored by: Prateek Singh**

Course: ALY6015 – Intermediate Analytics

Submission Date: 02/01/2022

**Instructor’s name: Jiyoung Yun**

# Introduction

The Chi-Square Test, also called as Pearson’s chi-squared test, is the result of data analysis based on observations of a set of random variables. The test is usually applied to comparing two statistical data sets. The chi-square test evaluates the likelihood of observations made by assuming the null hypothesis to be true. The formula for Chi-Square Test is displayed below:

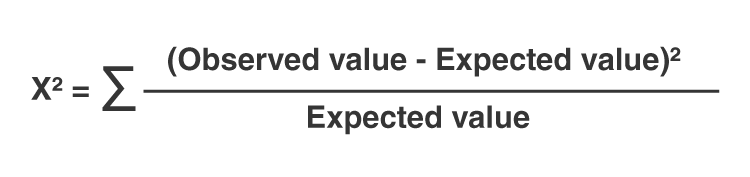


Fig. 1: Chi-Square Formula

ANOVA, or Analysis of Variance, is a statistical method for analyzing the difference between two or more groups mean based on the independent variable. A one-way ANOVA is performed with one independent variable, while a two-way ANOVA is performed with two independent variables.

This assignment is designed to utilize the understanding of chi-square and ANOVA techniques to efficiently process and analyze data or dataset to understand and summarize the trend.

# Analysis

Question 1:

There are four Blood Types given – Type A, B, O & AB. The expected distribution for all the blood types is given as 0.20, 0.28, 0.36 & 0.16 and the observed comes out to be 12, 8, 24 & 6 respectively from the random sample of 50 selected patients. The α is given as 0.10.

The Hypothesis is stated below:

H0: Type A = 0.20, Type B = 0.28, Type O = 0.36, Type AB = 0.16

H1: The distribution is not the same as stated in Null Hypothesis

The Critical Value is calculated to be 6.25. The chi-square test gives the p-value as 0.1404 which is greater than the α value (0.10) and hence we Fail To Reject Null Hypothesis, which means the distribution of population is correct and same as stated in the Null Hypothesis.

Question 2:

On-time performance by the airlines is categorized by the Bureau of Transportation Statistics as On-time, National Aviation System Delay, Aircraft arriving late and Other due to weather and other conditions. The percentage of time the above listed events happening is described as follows 0.708, 0.12, 0.082 & 0.09 and the observed comes out to be 125, 40, 10 & 25 respectively from the random sample of 200 selected flights. The α is given as 0.05.

The Hypothesis is stated below:

H0: The results from the government statistic are correct and valid

H1: The results differ from the government statistic

The Critical Value is calculated to be 7.81. The chi-square test gives the p-value as 0.0004763 which is lesser than the α value (0.05) and hence we Reject the Null Hypothesis, which means the results differ from the government statistic.

Question 3:

Numbers of admissions (in thousands) for two different years (2013 & 2014) is indicated. The number of admissions for different ethnicity is recorded as below table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Caucasian | Hispanic | African-American | Other |
| 2013 | 724 | 335 | 174 | 107 |
| 2014 | 370 | 292 | 152 | 140 |

The α is given as 0.05.

The Hypothesis is stated below:

H0: The movie attendance by year is independent of ethnicity

H1: The movie attendance by year is dependent of ethnicity

The Critical Value is calculated to be 7.81. The chi-square test gives the p-value as 5.478e-13 which is lesser than the α value (0.05) and hence we Reject the Null Hypothesis, which means the movie attendance by year is dependent of ethnicity.

Question 4:

Below table lists the numbers of officers and enlisted personnel for women in the military:

|  |  |  |
| --- | --- | --- |
| Action | Officers | Enlisted |
| Army | 10791 | 62491 |
| Navy | 7816 | 42750 |
| Marine Corps | 932 | 9525 |
| Air Force | 11819 | 54344 |

The α is given as 0.05.

The Hypothesis is stated below:

H0: The relationship exists between rank and branch of the Armed Forces

H1: The relationship doesn't exist between rank and branch of the Armed Forces

The Critical Value is calculated to be 7.81. The chi-square test gives the p-value as 1.726418e-141 which is lesser than the α value (0.05) and hence we Reject the Null Hypothesis, which means the relationship doesn't exist between rank and branch of the Armed Forces.

Question 5:

The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed below:

|  |  |  |
| --- | --- | --- |
| Condiments | Cereals | Desserts |
| 270 | 260 | 100 |
| 130 | 220 | 180 |
| 230 | 290 | 250 |
| 180 | 290 | 250 |
| 80 | 200 | 300 |
| 70 | 320 | 360 |
| 200 | 140 | 300 |
|  |  | 160 |

The α is given as 0.05.

The Hypothesis is stated below:

H0: µ1 = µ2 = µ3

H1: At least one mean out of three is different

The summary of ANOVA Test is displayed below. The F-value comes out to be 2.399 and the p-value as 0.118.

Text

Description automatically generated with medium confidence

Fig. 2: Anova Test Summary

The degrees of freedom for numerator and denominator is calculated as 2 and 19 respectively, and because of the fact that p-value is greater than alpha, we can safely conclude that we fail to reject the null Hypothesis which states the mean of Sodium amount in listed food types (Condiments, Cereals & Desserts) is equal.

However, if we were to reject Null Hypothesis if the sodium mean for even one food type was different, we will run TukeyHSD() function to check the difference in mean.

Text

Description automatically generated

Fig. 3: TukeyHSD Function Summary

Question 6:

The sales in millions of dollars for a year of a sample of leading companies are shown below:

|  |  |  |
| --- | --- | --- |
| Cereal | Chocolate | Candy Coffee |
| 578 | 311 | 261 |
| 320 | 106 | 185 |
| 264 | 109 | 302 |
| 249 | 125 | 689 |
| 237 | 173 |  |

The α is given as 0.01.

The Hypothesis is stated below:

H0: µ1 = µ2 = µ3

H1: At least one mean out of three is different

The summary of ANOVA Test is displayed below. The F-value comes out to be 2.172 and the p-value as 0.16.

Text

Description automatically generated with medium confidence

Fig. 4: Anova Test Summary

The degrees of freedom for numerator and denominator is calculated as 2 and 11 respectively, and because of the fact that p-value is greater than alpha, we can safely conclude that we fail to reject the null Hypothesis which states the sales mean in millions of dollars for a year of a sample of leading companies are same.

However, if we were to reject Null Hypothesis if the Sales mean for even one company sample was different, we will run TukeyHSD() function to check the difference in mean.

Text

Description automatically generated

Fig. 5: TukeyHSD Function Summary

Question 7:

The expenditures (in dollars) per pupil for states in three sections of the country are listed below:

|  |  |  |
| --- | --- | --- |
| Eastern Third | Middle Third | Western Third |
| 4946 | 6149 | 5282 |
| 5953 | 7451 | 8605 |
| 6202 | 6000 | 6528 |
| 7243 | 6479 | 6911 |
| 6113 |  |  |

The α is given as 0.05.

The Hypothesis is stated below:

H0: µ1 = µ2 = µ3

H1: At least one mean out of three is different

The summary of ANOVA Test is displayed below. The F-value comes out to be 0.649 and the p-value as 0.543.

A picture containing text

Description automatically generated

Fig. 6: Anova Test Summary

The degrees of freedom for numerator and denominator is calculated as 2 and 10 respectively, and because of the fact that p-value is greater than alpha, we can safely conclude that we fail to reject the null Hypothesis which states the expenditure mean in dollars per pupil for states in three sections of the country are same.

However, if we were to reject Null Hypothesis if the expenditure mean for even one section of the country was different, we will run TukeyHSD() function to check the difference in mean.

Text

Description automatically generated

Fig. 7: TukeyHSD Function Summary

Question 8:

A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a “Grow-light” in two different strengths and a plant food supplement with different mineral supplements. After several days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes below:

|  |  |  |
| --- | --- | --- |
|  | Grow-light 1 | Grow-light 2 |
| Plant Food A | 9.2, 9.4, 8.9 | 8.5, 9.2, 8.9 |
| Plant Food B | 7.1, 7.2, 8.5 | 5.5, 5.8, 7.6 |

The α is given as 0.05.

The Hypothesis is stated below:

H1: The mean of observations grouped by Grow-Light are same

H2: The mean of observations grouped by Plant Food are same

H3: There is no interaction between the Grow-Light and Plant Food

To start with we first need to create a Data Frame to store the values.

A picture containing chart

Description automatically generated

Fig. 8: Plant Measurement Data Frame

The summary of Two-way ANOVA Test is displayed below. The F-value for the two growth lights 1 and 2 have less difference as the F value, 3.6805, is low and the p-value is 0.09 more than alpha 0.05 and significant to 0.1 level. Inversely, the F-value for the two plant food type A and B have huge significant difference as the F value, 24.5623, is huge and the p-value is 0.001 is very small lesser than alpha 0.05 and significant to 0.01 level. Finally, the interaction of the two factors whose mean we are looking for, we can see that the F-statistic of 1.4377 is very low and the p-value of 0.2648 is greater than the alpha value. After concluding all the findings from the two-way ANOVA test we fail to reject the null hypothesis 1 and 3 and reject the null hypothesis 2. There is no difference in mean growth with respect to light. However, with respect to plant food, there is a difference in mean growth. Also, there is no interaction between the two factors studied.

Text

Description automatically generated

Fig. 9: Two-Way Anova Test Summary

The boxplot below also shows the difference in the mean of observations grouped by Plant Food types A and B.

Graphical user interface, application

Description automatically generated

Fig. 10: Boxplot for Mean growth vs Plant Food Type A & B

To conclude the test, because there is difference in mean of plant growth with respect to the food types, I conducted the TukeyHSD() test to find that difference. The output is displayed below.

Text

Description automatically generated

Fig. 11: TukeyHSD Function Summary

Question 9:

In this part, I imported the baseball dataset provided into the R. Post analysis and describing the dataset, we get the mean, standard deviation, median, skewness and kurtosis of all the variables.

Text

Description automatically generated

Fig. 12: Description of Baseball Dataset

To see the distribution of runs scored by all the teams throughout using the boxplot and arranging them with their median values, we can see team COL seems to be the highest performing team and team WSA is the lowest. There are few teams having some outliers as well.

A picture containing text, writing implement, stationary, pencil

Description automatically generated

Fig. 13: Distribution of Runs Scored

To follow on the Team’s wins distribution, I calculated the total wins of individual teams and plotted in the histogram. The distribution of wins by individual teams isn’t normal and is negatively skewed. The maximum number of wins is coming from the range between 3500 and 4000.

Chart, histogram

Description automatically generated

Fig. 14: Distribution of Wins

Further, we needed to analyze if there is a difference in the number of wins by decade. I created a decade variable in the main dataset and inked the respective decade for individual observations. Later we calculated the total number of wins grouped by the decades and plotted the histogram to visually identify if there is any relation in the distribution.

Text

Description automatically generated

Fig. 15: Decade Table

The histogram plot doesn’t conclude visually the normality in the distribution and the graph looks like negatively skewed. The mean and median are also calculated, and it comes around 16612.33 and 17953 respectively. The maximum wins were recorded in decade 2000, and the lowest in the decade 2010. Mathematically, we got the sense that the mean of wins by decade is not the same, however we are going to identify this using the hypothesis testing too.

Graphical user interface, application

Description automatically generated

Fig. 16: Distribution of Wins grouped by Decade

Assuming the expected frequencies are equal for wins by decade, the hypothesis stated is:

H0: There is no difference in the number of wins by decade

H1: There is a difference in the number of wins by decade

The α is given as 0.05.

The Critical Value is calculated to be 11.0705. The chi-square test gives the p-value as less than 2.2e-16 which is very low and lesser than the α value (0.05) and hence we Reject the Null Hypothesis, which means there is a difference in the number of wins by decade.

# Conclusion

1. We saw from the exercise that Chi-Square Test and One-way & Two-way ANOVA Test is utilized to analyze the hypothesis stated.
2. Baseball database hypothesis was analyzed using Chi-Square Goodness of Fitness Test to determine if there is a difference in the number of wins by decade and can be concluded in its favor.

# Reference

1. Admin. (2021, April 16). Chi-square test: How to calculate Chi-square using formula with example. BYJUS. Retrieved February 1, 2022, from <https://byjus.com/maths/chi-square-test/>
2. Bevans, R. (2021, January 7). An introduction to the one-way ANOVA. Scribbr. Retrieved February 1, 2022, from <https://www.scribbr.com/statistics/one-way-anova/>
3. Loughlin, E. O. (2021, June 7). How to... perform a two-way ANOVA in R #93. YouTube. Retrieved February 1, 2022, from <https://www.youtube.com/watch?v=7uR9c_yA0pE>
4. sgibb. (2013, October 15). How to set xlim in a hist( )plot while showing the full range of the variable in a histogram. Stack Overflow. Retrieved February 1, 2022, from <https://stackoverflow.com/questions/19375779/how-to-set-xlim-in-a-hist-plot-while-showing-the-full-range-of-the-variable-in>
5. Zach. (2021, June 14). How to convert character to factor in R (with examples). Statology. Retrieved February 1, 2022, from <https://www.statology.org/character-to-factor-in-r/>

# Appendix

# ALY6015 Module 2 R Practice: Singh Prateek ------------------------------------------------

#----------------- Author: Prateek Singh

#----------------- Submission Date: 25th Jan, 2022

#----------------- Tutor: Jiyoung Yun

# Step: Importing Libraries ------------------------------------------------

library(psych)

library(ggplot2)

library(ggpubr)

library(dplyr)

library(tibble)

# Question 1: ------------------------------------------------

# A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population.

# The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%.

# He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood.

# At α = 0.10, can it be concluded that the distribution is the same as that of the general population?

# Blood Type Expected | Observed

# Type A 20% | 12

# Type B 28% | 8

# Type O 36% | 24

# Type AB 16% | 6

# Stating the Hypothesis:

# H0: Type A = 0.20, Type B = 0.28, Type O = 0.36, Type AB = 0.16

# H1: The distribution is not the same as stated in Null Hypothesis

# df = (number of rows - 1) \* (number of columns -1)

# Critical Value = qchisq(alpha, df, lower.tail = F)

cvalue\_BloodType <- qchisq(0.10, 3, lower.tail = F)

cvalue\_BloodType

alpha\_1 <- 0.10

observed\_1 <- c(12, 8, 24, 6)

exp\_prob\_1 <- c(0.20, 0.28, 0.36, 0.16)

ChisqTest\_Result\_BloodType <- chisq.test(x = observed\_1, p = exp\_prob\_1)

ChisqTest\_Result\_BloodType$statistic

ChisqTest\_Result\_BloodType$p.value

ChisqTest\_Result\_BloodType$parameter

ChisqTest\_Result\_BloodType

ifelse(ChisqTest\_Result\_BloodType$p.value > alpha\_1, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Fail to reject null Hypothesis: The distribution of population is correct and same as stated in the Null Hypothesis as p-value comes out to be greater than alpha.

# Question 2: ------------------------------------------------

# According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

# Action | % of Time

# On time | 70.8

# National Aviation System delay | 8.2

# Aircraft arriving late | 9.0

# Other (because of weather and other conditions) | 12.0

# Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time;

# 40 were delayed because of weather, 10 because of a National Aviation System delay, and the rest because of arriving late.

# At α = 0.05, do these results differ from the government’s statistics?

# Stating the Hypothesis:

# H0: The results from the government statistic is correct and valid

# H1: The results differ from the government statistic

# df = (number of rows - 1) \* (number of columns -1)

# Critical Value = qchisq(alpha, df, lower.tail = F)

cvalue\_AirlinePerformace <- qchisq(0.05, 3, lower.tail = F)

cvalue\_AirlinePerformace

alpha\_2 <- 0.05

observed\_2 <- c(125, 40, 10, 25)

exp\_prob\_2 <- c(0.708, 0.12, 0.082, 0.09)

ChisqTest\_Result\_AirlinePerformace <- chisq.test(x = observed\_2, p = exp\_prob\_2)

ChisqTest\_Result\_AirlinePerformace$statistic

ChisqTest\_Result\_AirlinePerformace$p.value

ChisqTest\_Result\_AirlinePerformace$parameter

ChisqTest\_Result\_AirlinePerformace

ifelse(ChisqTest\_Result\_AirlinePerformace$p.value > alpha\_2, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Reject the null Hypothesis: The results differ from the government statistic as p-value comes out to be lesser than alpha.

# Question 3: ------------------------------------------------

# Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years.

# At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

# Caucasian | Hispanic | African American | Other

# 2013 724 | 335 | 174 | 107

# 2014 370 | 292 | 152 | 140

# Stating the Hypothesis:

# H0: The movie attendance by year is independent of ethnicity

# H1: The movie attendance by year is dependent of ethnicity

# df = (number of rows - 1) \* (number of columns -1)

# Critical Value = qchisq(alpha, df, lower.tail = F)

cvalue\_EthnicityAttendance <- qchisq(0.05, 3, lower.tail = F)

cvalue\_EthnicityAttendance

alpha\_3 <- 0.05

Yr2013 <- c(724, 335, 174, 107)

Yr2014 <- c(370, 292, 152, 140)

MovieAttendence\_Matrix <- matrix(c(Yr2013, Yr2014), nrow = 2, byrow = TRUE,

dimnames = list(c("2013", "2014"), c("Caucasian", "Hispanic", "African American", "Other")))

ChisqTest\_Result\_EthnicityAttendance <- chisq.test(MovieAttendence\_Matrix)

ChisqTest\_Result\_EthnicityAttendance$statistic

ChisqTest\_Result\_EthnicityAttendance$p.value

ChisqTest\_Result\_EthnicityAttendance$parameter

ChisqTest\_Result\_EthnicityAttendance

ifelse(ChisqTest\_Result\_EthnicityAttendance$p.value > alpha\_3, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Reject the null Hypothesis: The movie attendance by year is dependent of ethnicity as p-value comes out to be lesser than alpha.

# Question 4: ------------------------------------------------

# This table lists the numbers of officers and enlisted personnel for women in the military.

# At α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

# Action Officers | Enlisted

# Army 10,791 | 62,491

# Navy 7,816 | 42,750

# Marine Corps 932 | 9,525

# Air Force 11,819 | 54,344

# Stating the Hypothesis:

# H0: The relationship exists between rank and branch of the Armed Forces

# H1: The relationship doesn't exist between rank and branch of the Armed Forces

# df = (number of rows - 1) \* (number of columns -1)

# Critical Value = qchisq(alpha, df, lower.tail = F)

cvalue\_ArmedForces <- qchisq(0.05, 3, lower.tail = F)

cvalue\_ArmedForces

alpha\_4 <- 0.05

Army <- c(10791, 62491)

Navy <- c(7816, 42750)

MarineCorps <- c(932, 9525)

AirForce <- c(11819, 54344)

ArmedForces\_Matrix <- matrix(c(Army, Navy, MarineCorps, AirForce), nrow = 4, byrow = TRUE,

dimnames = list(c("Army", "Navy", "Marine Corps", "Air Force"), c("Officers", "Enlisted")))

ChisqTest\_Result\_ArmedForces <- chisq.test(ArmedForces\_Matrix)

ChisqTest\_Result\_ArmedForces$statistic

ChisqTest\_Result\_ArmedForces$p.value

ChisqTest\_Result\_ArmedForces$parameter

ChisqTest\_Result\_ArmedForces

ifelse(ChisqTest\_Result\_ArmedForces$p.value > alpha\_4, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Reject the null Hypothesis: The relationship doesn't exist between rank and branch of the Armed Forces as p-value comes out to be lesser than alpha.

# One-Way ANOVA

# Question 5: ------------------------------------------------

# The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed.

# At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

# Condiments | Cereals | Desserts

# 270 | 260 | 100

# 130 | 220 | 180

# 230 | 290 | 250

# 180 | 290 | 250

# 80 | 200 | 300

# 70 | 320 | 360

# 200 | 140 | 300

# | | 160

# Stating the Hypothesis:

# H0: µ1 = µ2 = µ3

# H1: At least one mean out of three is different

alpha\_5 <- 0.05

Condiments <- data.frame('sodium' = c(270, 130, 230, 180, 80, 70, 200), 'food' = rep('Condiments', 7), stringsAsFactors = FALSE)

Cereals <- data.frame('sodium' = c(260, 220, 290, 290, 200, 320, 140), 'food' = rep('Cereals', 7), stringsAsFactors = FALSE)

Desserts <- data.frame('sodium' = c(100, 180, 250, 250, 300, 360, 300, 160), 'food' = rep('Desserts', 8), stringsAsFactors = FALSE)

Sodium <- rbind(Condiments, Cereals, Desserts)

Sodium$food <- as.factor(Sodium$food)

Anova\_Food <- aov(sodium ~ food, data = Sodium)

summary(Anova\_Food)

Anova\_Food\_Summary <- summary(Anova\_Food) # Save summary to an object

# Degrees of freedom

# K-1: between group variance: Numerator

df.numerator <- Anova\_Food\_Summary[[1]][1, "Df"]

df.numerator

# N-K: within group variance: Denominator

df.denominator <- Anova\_Food\_Summary[[1]][2, "Df"]

df.denominator

F.value <- Anova\_Food\_Summary[[1]][1, "F value"] # Extract F test value

F.value

p.value <- Anova\_Food\_Summary[[1]][1, "Pr(>F)"] # Extract p-value

p.value

ifelse(p.value > alpha\_5, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Fail to reject the null Hypothesis: The mean of Sodium amount in listed food types(Condiments, Cereals & Desserts) is equal as p-value comes out to be greater than alpha.

# However, if we were to reject Null Hypothesis if the sodium mean for even one food type was different, we will run TukeyHSD() function to check the difference in mean

TukeyHSD(Anova\_Food)

# Question 6: ------------------------------------------------

# The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means?

# Cereal | Chocolate | Candy Coffee

# 578 | 311 | 261

# 320 | 106 | 185

# 264 | 109 | 302

# 249 | 125 | 689

# 237 | 173

# Stating the Hypothesis:

# H0: µ1 = µ2 = µ3

# H1: At least one mean out of three is different

alpha\_6 <- 0.01

Cereal <- data.frame('Sales' = c(578, 320, 264, 249, 237), 'companies' = rep('Cereal', 5), stringsAsFactors = FALSE)

Chocolate <- data.frame('Sales' = c(311, 106, 109, 125, 173), 'companies' = rep('Chocolate', 5), stringsAsFactors = FALSE)

Candy\_Coffee <- data.frame('Sales' = c(261, 185, 302, 689), 'companies' = rep('Candy\_Coffee', 4), stringsAsFactors = FALSE)

Sales <- rbind(Cereal, Chocolate, Candy\_Coffee)

Sales$companies <- as.factor(Sales$companies)

Anova\_Companies <- aov(Sales ~ companies, data = Sales)

summary(Anova\_Companies)

Anova\_Companies\_Summary <- summary(Anova\_Companies) # Save summary to an object

# Degrees of freedom

# K-1: between group variance: Numerator

df.numerator\_Sales <- Anova\_Companies\_Summary[[1]][1, "Df"]

df.numerator\_Sales

# N-K: within group variance: Denominator

df.denominator\_Sales <- Anova\_Companies\_Summary[[1]][2, "Df"]

df.denominator\_Sales

F.value\_Sales <- Anova\_Companies\_Summary[[1]][1, "F value"] # Extract F test value

F.value\_Sales

p.value\_Sales <- Anova\_Companies\_Summary[[1]][1, "Pr(>F)"] # Extract p-value

p.value\_Sales

ifelse(p.value\_Sales > alpha\_6, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Fail to reject the null Hypothesis: The sales mean in millions of dollars for a year of a sample of leading companies are same as p-value comes out to be greater than alpha.

# However, if we were to reject Null Hypothesis if the Sales mean for even one company sample was different, we will run TukeyHSD() function to check the difference in mean

TukeyHSD(Anova\_Companies)

# Question 7: ------------------------------------------------

# The expenditures (in dollars) per pupil for states in three sections of the country are listed.

# Using α = 0.05, can you conclude that there is a difference in means?

# Eastern third | Middle third | Western third

# 4946 | 6149 | 5282

# 5953 | 7451 | 8605

# 6202 | 6000 | 6528

# 7243 | 6479 | 6911

# 6113 |

# Stating the Hypothesis:

# H0: µ1 = µ2 = µ3

# H1: At least one mean out of three is different

alpha\_7 <- 0.05

Eastern\_Third <- data.frame('Expenditure' = c(4946, 5953, 6202, 7243, 6113), 'Sections' = rep('Eastern Third', 5), stringsAsFactors = FALSE)

Middle\_Third <- data.frame('Expenditure' = c(6149, 7451, 6000, 6479), 'Sections' = rep('Middle Third', 4), stringsAsFactors = FALSE)

Western\_Third <- data.frame('Expenditure' = c(5282, 8605, 6528, 6911), 'Sections' = rep('Western Third', 4), stringsAsFactors = FALSE)

Expenditure <- rbind(Eastern\_Third, Middle\_Third, Western\_Third)

Expenditure$Sections <- as.factor(Expenditure$Sections)

Anova\_Sections <- aov(Expenditure ~ Sections, data = Expenditure)

summary(Anova\_Sections)

Anova\_Sections\_Summary <- summary(Anova\_Sections) # Save summary to an object

# Degrees of freedom

# K-1: between group variance: Numerator

df.numerator\_Exp <- Anova\_Sections\_Summary[[1]][1, "Df"]

df.numerator\_Exp

# N-K: within group variance: Denominator

df.denominator\_Exp <- Anova\_Sections\_Summary[[1]][2, "Df"]

df.denominator\_Exp

F.value\_Exp <- Anova\_Sections\_Summary[[1]][1, "F value"] # Extract F test value

F.value\_Exp

p.value\_Exp <- Anova\_Sections\_Summary[[1]][1, "Pr(>F)"] # Extract p-value

p.value\_Exp

ifelse(p.value\_Exp > alpha\_7, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Fail to reject the null Hypothesis: The expenditure mean in dollars per pupil for states in three sections of the country are same as p-value comes out to be greater than alpha.

# However, if we were to reject Null Hypothesis if the expenditure mean for even one section of the country was different, we will run TukeyHSD() function to check the difference in mean

TukeyHSD(Anova\_Sections)

# Question 8: ------------------------------------------------

# A gardening company is testing new ways to improve plant growth.

# Twelve plants are randomly selected and exposed to a combination of two factors, a “Grow-light” in two different strengths and a plant food supplement with different mineral supplements.

# After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes.

# | Grow-light 1 | Grow-light 2

# Plant Food A | 9.2, 9.4, 8.9 | 8.5, 9.2, 8.9

# Plant Food B | 7.1, 7.2, 8.5 | 5.5, 5.8, 7.6

# Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food?

# Use α = 0.05.

# Stating the Hypothesis:

# H1: The mean of observations grouped by Grow-Light are same

# H2: The mean of observations grouped by Plant Food are same

# H3: There is no interaction between the Grow-Light and Plant Food

alpha\_8 <- 0.05

Plant\_Measure = data.frame(PlantFood = rep(c("A", "B"), each=6),

GrowLight=rep(c("1", "2"), each=3, times=2),

Growth=c(9.2,9.4,8.9,8.5,9.2,8.9,7.1,7.2,8.5,5.5,5.8,7.6))

Plant\_Measure

TwoWayAnova\_PlantGrowth <- aov(Growth ~ GrowLight \* factor(PlantFood), data = Plant\_Measure)

anova(TwoWayAnova\_PlantGrowth)

plot(Growth ~ GrowLight + factor(PlantFood), data = Plant\_Measure, xlab = "Plant Food")

TukeyHSD(TwoWayAnova\_PlantGrowth)

# Question 9: ------------------------------------------------

Baseball <- read.csv('baseball.csv')

describe(Baseball)

Baseball <- as.data.frame(unclass(Baseball), stringsAsFactors = TRUE)

ggplot(Baseball, aes(x = reorder(Team, RS, FUN = median), y = RS, fill = Team)) +

geom\_boxplot(notch = FALSE, alpha = 0.9) +

scale\_y\_continuous(name = "Run Scored", breaks = c(0, Baseball$RS), guide = guide\_axis(check.overlap = TRUE)) +

coord\_flip() +

stat\_boxplot(geom = 'errorbar', width = 0.5) +

scale\_x\_discrete(name = "Team") +

labs(title = "Teams Runs Scored") +

theme(plot.title = element\_text(hjust = 0.5, color = "red", face = "bold"),

axis.title.x = element\_text(color = "Orange"),

axis.title.y = element\_text(color = "orange"),

axis.text.x = element\_text(color = "cyan4", angle = 90),

axis.text.y = element\_text(color = "cyan4"),

axis.line = element\_line(color = "cyan4")

)

wins <- Baseball %>%

group\_by(Team) %>%

summarize(wins = sum(W)) %>%

as\_tibble()

hist(wins$wins, main = "Distribution of Team's Wins", xlab = "Number of Wins Per Team", col = "Pink")

Baseball$Decade <- Baseball$Year - (Baseball$Year %% 10)

Decade\_Win <- Baseball %>%

group\_by(Decade) %>%

summarize(wins = sum(W)) %>%

as\_tibble()

Decade\_Win

hist(Decade\_Win$wins, main = "Distribution of Wins per Decade", xlab = "Number of Wins Per Decade", xaxt = "n")

axis(1, at = seq(5000, 25000, by = 2000), labels = seq(5000, 25000, by = 2000))

mean(Decade\_Win$wins)

median(Decade\_Win$wins)

# Assuming the expected frequencies are equal for wins by decade

# Stating the Hypothesis:

# H0: There is no difference in the number of wins by decade

# H1: There is a difference in the number of wins by decade

# df = (number of rows - 1) \* (number of columns -1)

# Critical Value = qchisq(alpha, df, lower.tail = F)

cvalue\_WinsByDecade <- qchisq(0.05, 5, lower.tail = F)

cvalue\_WinsByDecade

alpha\_9 <- 0.05

ChisqTest\_Result\_WinsByDecade <- chisq.test (Decade\_Win)

ChisqTest\_Result\_WinsByDecade$statistic

ChisqTest\_Result\_WinsByDecade$p.value

ChisqTest\_Result\_WinsByDecade$parameter

ChisqTest\_Result\_WinsByDecade

ifelse(ChisqTest\_Result\_WinsByDecade$p.value > alpha\_9, "Fail to reject null hypothesis", "Reject the null hypothesis")

# Reject null Hypothesis: There is a difference in the number of wins by decade as p-value comes out to be greater than alpha.