

*Toronto*

**Hypothesis Testing: Chi-square and ANOVA**

Soni Manana

a *College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6015* ***NUID****: 002982645*

Under the guidance of

**Dr. Prof. Alex Maizlish**

**Introduction**

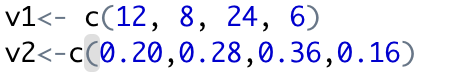
We have provided with questions with the learning purpose which are of hypothesis testing using Chi-square and ANOVA tests.

**About Chi-Square and ANOVA**

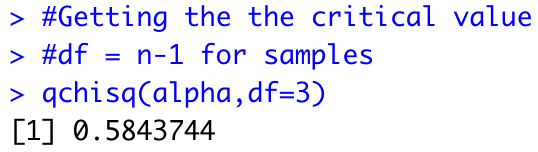
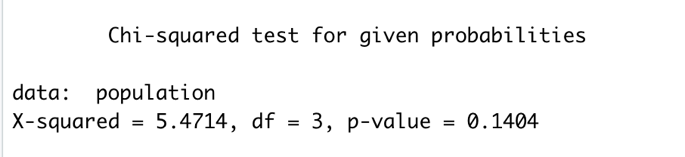
* A Chi-square test is a hypothesis testing. Chi-square tests involves p-value, lower bound and upper bound and X2 value to judge or determine significance. If all the variables are categorical we can’t use ANOVA but Chi-Square.
* The less X2 squared value is, more correlation. They have inverse relationship.
* **ANOVA** stands for **Analysis of Variance** and to be specific, used to determine significance between variables ranges from one, two or three. If all the variables have numerical and you need to check between difference in significance like making interaction, ANOVA is used.

**11-1 (A) Blood Types**

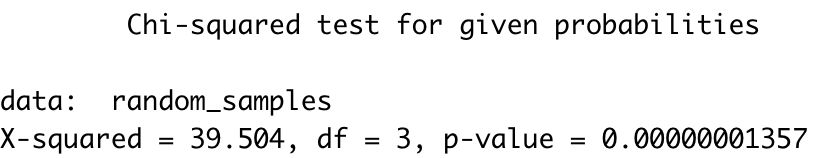
* Distribution of blood types for the general populations provided and they are type A, type B, Type O, type AB respective 20%, 28%, 36% and 16% and taking random samples from that population of 50 to determine and they got 12 having type A blood, 8 having type B, 24 having type O, and 6 having type AB blood.
* Level of significance has provided which is 0.1.
* **H0**: Random sample distributions are same as the general population’s.
* **H1:** Random sample distributions are not same as the general population’s.
  + Two vectors for test. Converted percentage into values.



* From the Chi-square test, p-value is 0.14 which is greater than 0.1. So we don’t have enough evidence to reject. In addition, X2 statistic value is greater than critical value which is 0.584. So we can’t reject **H0.**
* This **qchisq()** method does same as we manually find value from table.



**11-1 (B) On-Time Performance By Airlines**

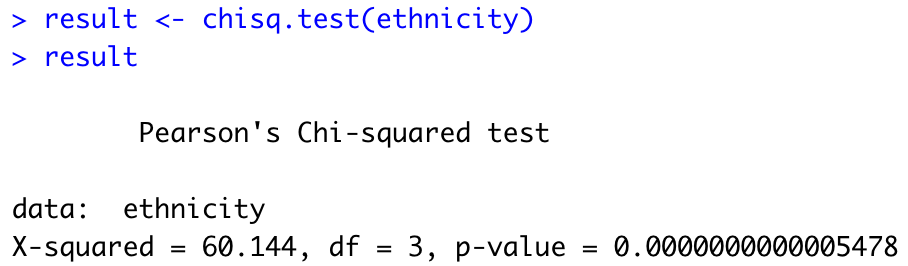
* Statistics of flights from airline and their punctuality was provided by the Bureau of Transportation Statistics. Probability of flight like whether it will be on-time or some kind of delay will be there. So 70.8% on-time departure, 8.2% delay because of National Aviation System delay, 9% Late arrival at the airport and 12% due to other reasons.
* To test, 200 samples has taken in which 125 of samples were on time, 40 were delayed because of some reason (other), 10 because of National Aviation System delay.
* We have to find for 95% level of significance.
* **H0**: The provided probabilities of flights and results of samples are same.
* **H1:** The provided probabilities of flights and results of samples aren’t same.
* 
* Note that p-value is just 0 which is less than alpha (0.05). So, we reject the null hypothesis.
* Critical value for this test is 0.3518 which is less than X2.

****

**11-2 (A) Ethnicity and Movie Admissions**

* Content of ethnicity is provided for two years which are 2013 and 2014 for movie admissions like Caucasian, Hispanic, African, etc.
* For this, 95% level of significance was provided.
* **H0**: Movie admissions are independent on ethnicity.
* **H1**: Movie admissions are not free of ethnicity.
* There is Table of movie admissions and ethnicity.

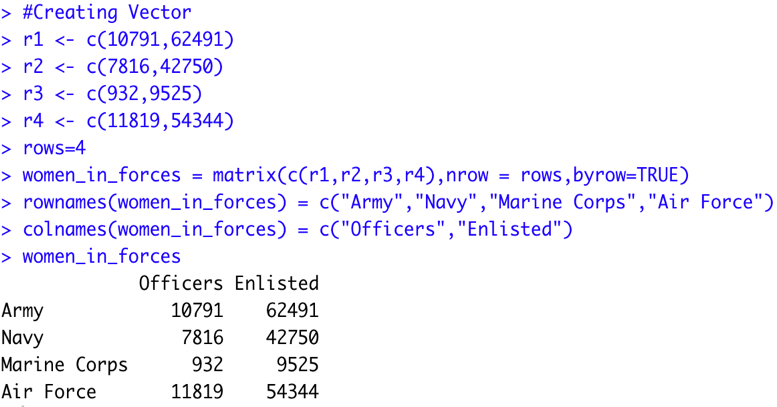


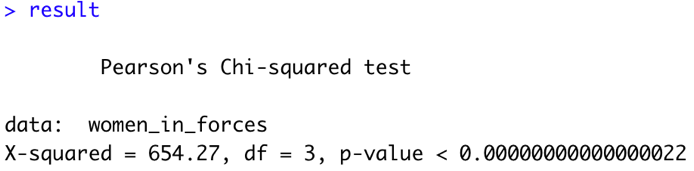


* Note that p-value is 0 with X2  value and Critical value is 7.814.
* Value of p is less than alpha and X2 is greater than critical value. So for both evidence, we reject the null hypothesis.

**11-2 (B) Women in Military**

* Table of data containing women in military has provided. For different, Arm forces like Army and Navy, and ranks are provided
* First needs to create a matrix. Then provide row names, column names according to table.
* All of the data is of women no other than that.
* **H0**: Ranks of officers have no relation with different arm forces.
* **H1:** There is dependency between rank and branch of forces.
* We need to check that at 95% level of significance which is 0.05 for statistically (alpha).

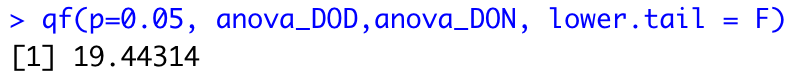
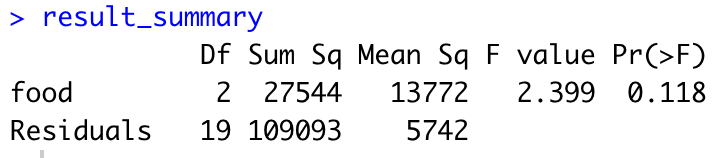




* Note that X2 value is high due to number but it is greater than critical value and p-value is also 0 which is less than value of alpha. In that case, we reject the null hypothesis and we found there is a relationship exists.

**12-1 Sodium Content of Food**

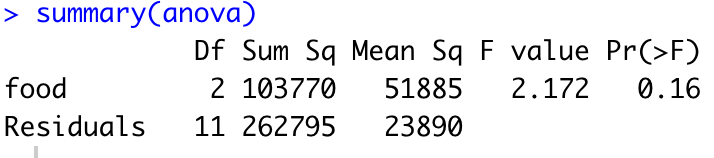
* Samples have been given in which amount of sodium in one serving of 3 different foods Items like Condiments, Cereals and Desserts.
* **H0**: The mean sodium difference in three of them are same. (u1 = u2 = u3)
* **H1:** Anyone of them has different mean of sodium.
* We need to find for 95% level of significance.
* We will use **ANOVA** for this problem as problem have numeric for 3 categories. By mean, Not all variables are categorical.



* Value of F is 2.399 and p is 0.118 which is less than 0.05.
* So we reject the null hypothesis.
* The high F-value in ANOVA, the high variance of sample with respect to variation within sample.

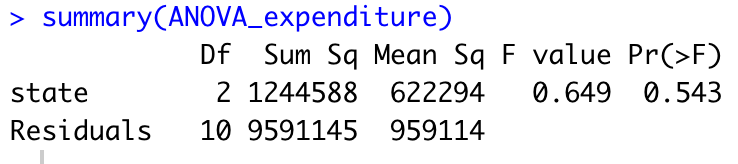
**12-2 (A) Sales for Leading Companies**

* The sales in millions of dollars for a specific year of leading companies are given. At 99% the level of significance to determine if there is difference in means are same.
* **H0:** The mean difference in the sales of products is same.
* **H1:** The mean difference in the sales of products is not same.
* F-value is 2.17 and value of p is 0.16 which greater than 0.01.
* So, we cannot reject the null hypothesis.
* One way is used to identify if one variable affects a response variable or not.



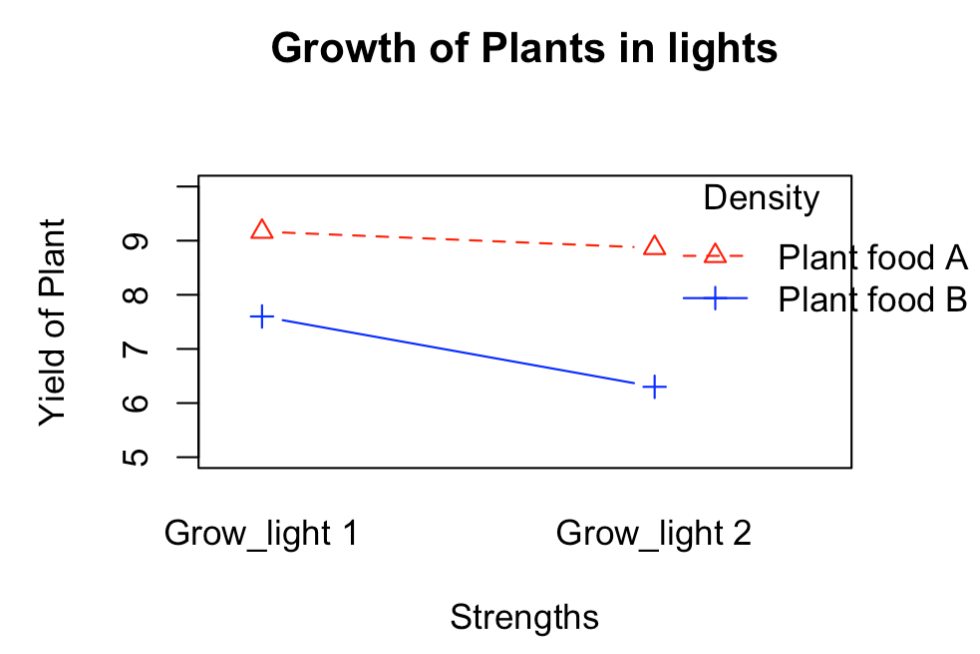
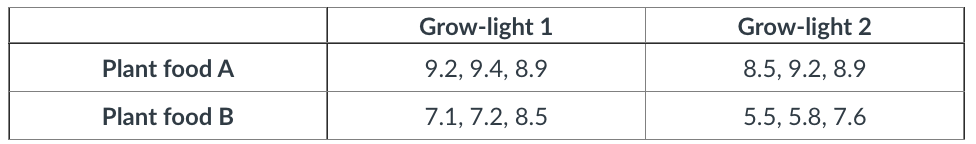
**12-2 (B) Per-Pupil Expenditures**

* Expenditures per pupils of states of three sections of country have given.
* We need to determine at 95% level of significance.
* **H0:** In all of the parts means of expenditures are same.
* **H1:** In all of the parts means of expenditures are not same. (μ1 != μ2 != μ3)
* For this test F-value is low which explains less variation and value of p is 0.543 which greater than alpha.
* Due to no evidence, we cannot reject the null hypothesis.

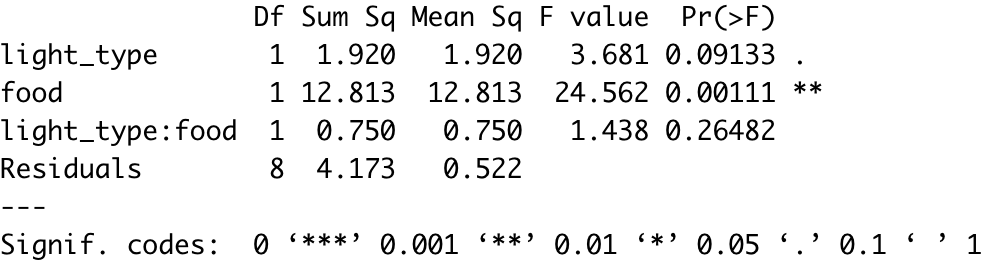


**12-3**

* A table of experiments conducted on one plant. It shows numbers regarding one plant can grow in two different strengths of grow-light and plant foods (A and B) supplement by a gardening company.



* Graph we can see that growth is good when there is interaction of Grow-light 1 and Plant food A where as comparing to that Grow-Light 2 slightly less effective impact on growth while using Plant food A. Plant food B has some kind of problem which has not good impact on growth as Plant food A in both the strenths of light
* Interaction has the value of p 0.26482 which is greater than the alpha. So we can’t reject the hypothesis.

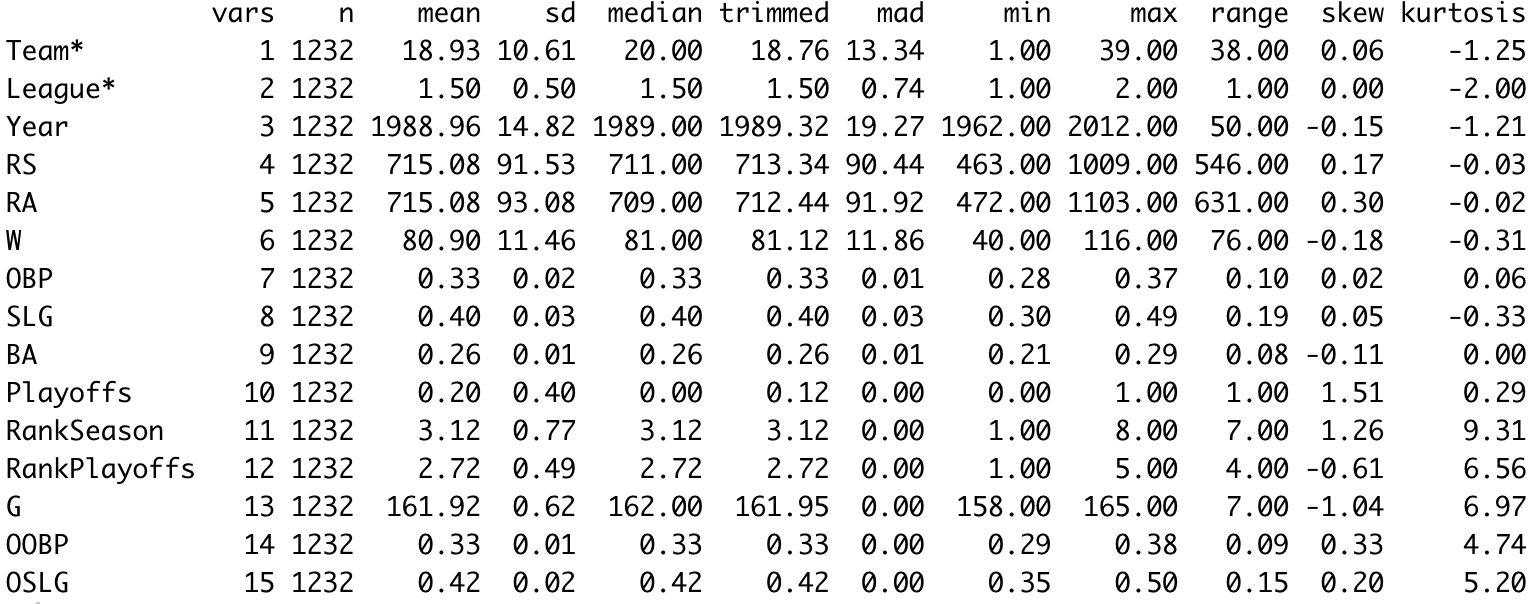
****

* Lowest value of F can also notice for the interaction.

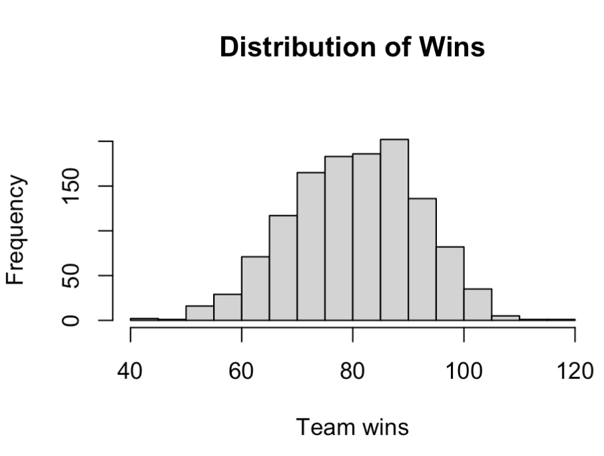
**Analysis of Baseball Game**

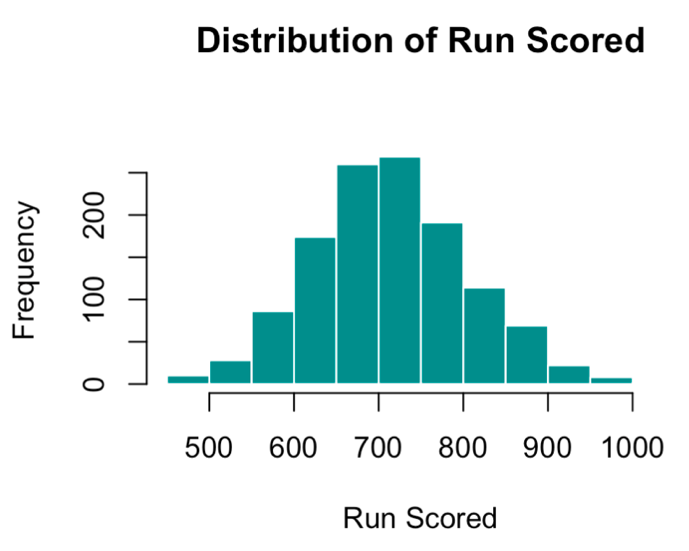
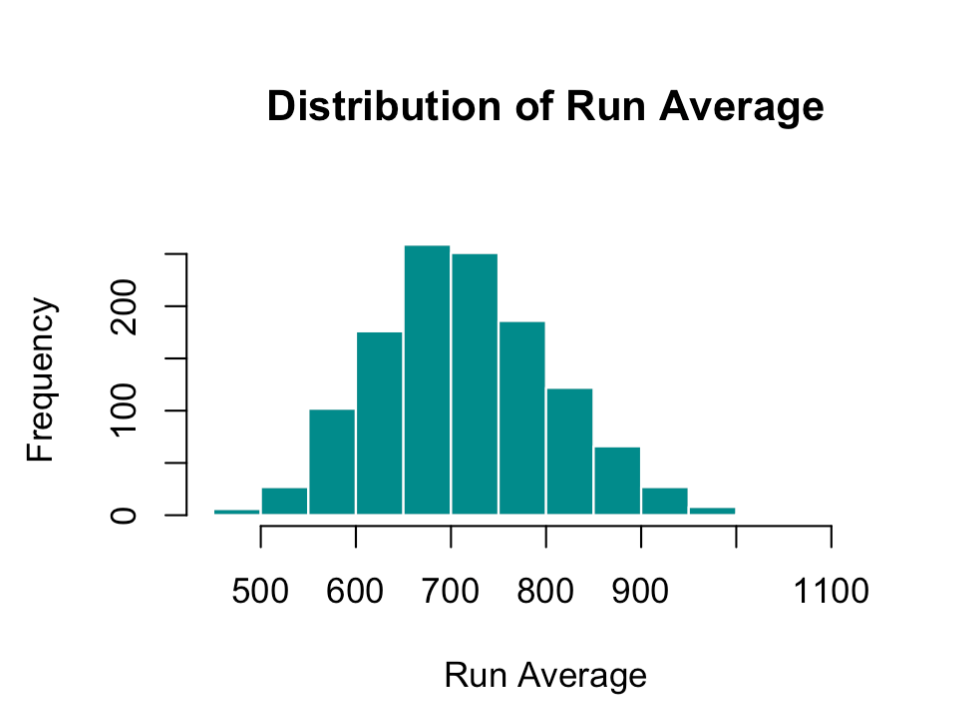
1. **Descriptive Analysis**

* Dataset was totally uncleaned, Approx. 70% of samples were affected with null values. So replaced mean instead removing rows. Total row count is 1232 with 15 columns.
* Note that Skewness is almost 0 for all variables which is good for normality check for all kinds of tests and analysis.

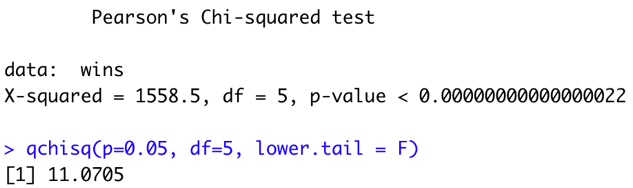
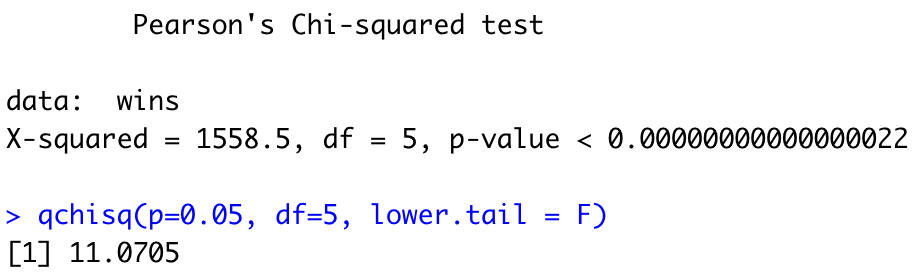
****

* From the histograms we can see that Run scored and Run Average both are around 700. However, there are some samples with run average over 1000.
* Run scored and run Average have almost same mean with same standard deviation so we can predict from mid of the match if how much score can be expected.
* Mean of other variables are just around 0.

****

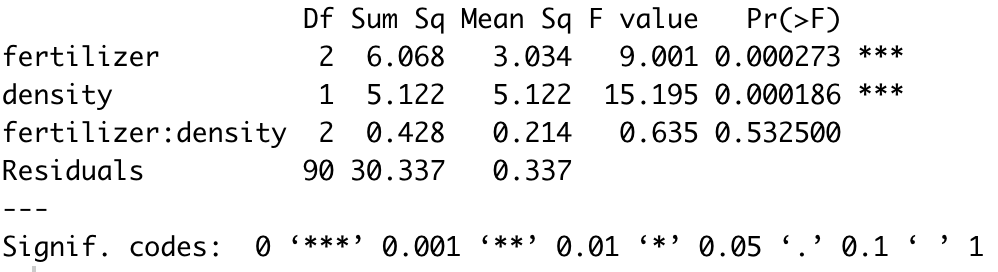
****

1. **Hypothesis testing of Wins with Decade**

* Data of Count of wins with Decades. If we can determine if there is something driving between two variables.
* For this test, level of significance Is 95%.
* To test, we need to state hypothesis, Therefore,
* **H0**: Count of win has some existing relation with decades.
* **H1**: Count of win is independent from time.
* 
* 
* Note that value of p is almost 0 which is less than the value of significance 0.05 and X2 statistics is 1558 which is far greater than 11.0705 which is critical value of test.
* So we have enough evidence to reject the null hypothesis.

**Hypothesis testing of Crop Data**

* After importing dataset, I’ve generated a data frame for it.
* Data has 4 variables like Block, Fertilizer, Yield and Density of crop.
* We need to identify if fertilizer and density have an impact on yield of crops.
* This will be two factor ANOVA test as specific interaction is required to test.
* **H0**: There is no effect on yield of crop from other factors like fertilizer, density and their interaction.
* **H1**: There is effect on yield of crop from the usage of fertilizer, density and their interaction.
* As from the P values the interaction has value 0.532. So we can’t reject the null hypothesis.
* So for interaction we can reject H0, but for other two independent, we can reject as p-values are 0. As we can note F-values decreased to 0 when they interacted.

****

**References**

Statology, *2022.* One-way vs Two-way ANOVA with: When to Use Each *Source:* https://www.statology.org/one-way-vs-two-way-anova/

Wikipedia, *2022. Interleague play Source:* <https://en.wikipedia.org/wiki/Interleague_play>

Scribbr, Rebecca Bevans, PhD 2022. Two-way ANOVA | When and How to Use it, With Examples *Source:* https://www.scribbr.com/statistics/two-way-anova/

Littleballparks, 2022. What does RA, RS MEAN in Baseball *Source:* https://littleballparks.com/what-does-ra-rd-and-rs-mean-in-baseball/

**Appendix**

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

library(DescTools)

library(dplyr)

library(tidyverse)

library(psych)

library(ggplot2)

# 11.1 (A)

#Creating the vector values of result of samples

v1<- c(12, 8, 24, 6)

v2<-c(0.20,0.28,0.36,0.16)

#Creating vector of distribution of general

#Result of chi square test

?chisq.test()

test\_result=chisq.test(x=random\_samples, p=general\_population)

test\_result

#Getting the the critical value

#df = n-1 for samples

qchisq(alpha,df=3)

t.test(p,population,conf.level=0.90)

# 11.1 (B)

#Significance value

alpha<- 0.05

#Creating the vector values.

random\_samples<- c(125, 40, 10, 25)

#Creating vector probabilities

flight\_probs<-c(0.708,0.082,0.090,0.12)

#Result of chi square test

test\_result<-chisq.test(x=random\_samples, p=flight\_probs)

test\_result

#Findind the critical value

qchisq(0.05,df=3)

# 11.2 (A)

#Creating vector

r1 <- c(724,335,174,107)

r2 <- c(370,292,152,140)

rows=2

ethnicity = matrix(c(r1,r2),nrow = rows,ncol=4,byrow=TRUE)

rownames(ethnicity) = c(2013,2014)

colnames(ethnicity) = c("Caucasian","Hispanic","African American","Other")

ethnicity

result <- chisq.test(ethnicity)

#Summary results

result$statistic

result$p.value

result$parameter

result$statistic

#Calculation critical value

qchisq(p=.05, df=3, lower.tail=FALSE)

# 11.2 (B)

#Creating Vector

r1 <- c(10791,62491)

r2 <- c(7816,42750)

r3 <- c(932,9525)

r4 <- c(11819,54344)

rows=4

women\_in\_forces = matrix(c(r1,r2,r3,r4),nrow = rows,byrow=TRUE)

rownames(women\_in\_forces) = c("Army","Navy","Marine Corps","Air Force")

colnames(women\_in\_forces) = c("Officers","Enlisted")

women\_in\_forces

result <- chisq.test(women\_in\_forces)

result

#Summary results

result$statistic

result$p.value

result$parameter

result$statistic

#Calculating critical value

a <- qchisq(p=.05, df=3, lower.tail=FALSE)

# 12.1

condiments<- data.frame('sodium' = c(270, 130, 230, 180, 80, 70, 200), 'type' = rep('condiments', 7), stringsAsFactors =F)

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

desserts <- data.frame('sodium' = c(100, 180, 250, 250, 300, 360, 300, 160), 'type'= rep('desserts', 8), stringsAsFactors =F)

value\_sodium<-rbind(condiments, cereals, desserts)

value\_sodium$type<- as.factor(sodium$type)

value\_sodium

#probs

p<-c(0.708,0.082,0.090,0.12)

#ANOVA test

s.anova<- aov(value\_sodium~food, data=value\_sodium)

smmry<-summary(s.anova)

smmry

# k - 1

nume <- a.summary[[1]][1, "Df"]

# N - k

domi <- a.summary[[1]][2, "Df"]

domi

#Finding critical value

qf(p=0.05, domi,nume, lower.tail = F)

# 12-2 (A)

alpha = 0.01

cereal <- data.frame('sales'=c(578,320,264,249,237),'food' = rep('cereal',5),stringsAsFactors = FALSE)

chocolate\_candy<- data.frame('sales'=c(311,106,109,125,173),'food'=rep('chocolate\_candy',5),stringsAsFactors = FALSE)

coffee <- data.frame('sales'=c(261,185,302,689),'food'=rep('coffee',4),stringsAsFactors = FALSE)

qf(p=.01, df1=2, df2=11, lower.tail=FALSE)

sales <- rbind(cereal,chocolate\_candy,coffee)

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

anova <- aov(sales~food,data = sales)

summary(anova)

# 12.2.2 (B)

alpha = 0.05

Eastern\_third<- data.frame('expenditure'=c(4946,5953,6202,7243,6113),'state' = rep('Eastern third',5),stringsAsFactors = FALSE)

Middle\_third<- data.frame('expenditure'=c(6149,7451,6000,6479),'state'=rep('Middle third',4),stringsAsFactors = FALSE)

Western\_third<- data.frame('expenditure'=c(5282,8605,6528,6911),'state'=rep('Western third',4),stringsAsFactors = FALSE)

expenditure <- rbind(Eastern\_third,Middle\_third,Western\_third)

expenditure$state <- as.factor(expenditure$state)

anova1 <- aov(expenditure~state,data = expenditure)

ANOVA\_expenditure = anova1

summary(ANOVA\_expenditure)

# 12.3

#Creating dataframe

plantA1<- data.frame('growth'=c(9.2,9.4,8.9),'light\_type' = rep('Grow\_light 1',3),'food'=rep("Plant food A",3),stringsAsFactors = FALSE)

plantA2<- data.frame('growth'=c(8.5,9.2,8.9),'light\_type' = rep('Grow\_light 2',3),'food'=rep("Plant food A",3),stringsAsFactors = FALSE)

plantB1<- data.frame('growth'=c(7.1,7.2,8.5),'light\_type' = rep('Grow\_light 1',3),'food'=rep("Plant food B",3),stringsAsFactors = FALSE)

plantB2<- data.frame('growth'=c(5.5,5.8,7.6),'light\_type' = rep('Grow\_light 2',3),'food'=rep("Plant food B",3),stringsAsFactors = FALSE)

company <- rbind(plantA1,plantA2,plantB1,plantB2)

company$light\_type <- as.factor(company$light\_type)

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

#Visualization

interaction.plot(company$light\_type , company$food, company$growth ,

type = "b" , col = c("red","blue"),xlab='Strengths',trace.label = "Density",

pch = c(2,3) ,ylab = "Yield of Plant", ylim = c(5,10),

main='Growth of Plants in lights')

#Applying ANOVA

anova1 <- aov(growth~light\_type+food+light\_type:food,data=company)

anova.summary <-summary(anova1)

anova.summary

# ON YOUR OWN #

#Importing dataset

df<-read.csv("baseball.csv")

summary(df)

str(df)

psych::describe(df)

# MEAN instead NA

for(i in 1:ncol(df)){

df[is.na(df[,i]), i] <- mean(df[,i], na.rm = TRUE)

}

# SUMMARIES

summary(df)

psych::describe(df)

win <- df %>% group\_by(Team) %>% summarize(wins = sum(W)) %>% as.tibble()

win\_by\_league <- df %>%

group\_by(League) %>% summarize(wins = sum(W)) %>%

as.tibble()

hist(df$W,main="Distribution of Winings",xlab='Team wins',col='cyan4',border='white')

hist(df$RS,main="Distribution of Run Scored",xlab='Run Scored',col='cyan4',border='white')

hist(df$RA,main="Distribution of Run Average",xlab='Run Average',col='cyan4',border='white')

# Extract decade from year

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

# Create a wins table by summing the wins by decade

wins <- df %>%

group\_by(Decade) %>% summarize(wins = sum(W)) %>%

as.tibble()

#Creating the vector values.

r1<- c(125, 40, 10, 25)

#Creating vector probabilities

p<-c(0.708,0.082,0.090,0.12)

#Result of chi square test

outcome<-chisq.test(wins)

outcome

qchisq(p=0.05, df=5, lower.tail = F)

crop <- read.csv("crop\_data.csv")

sum(is.na(crop))

#Data cleaning and preparing

crop$density <- as.factor(crop$density)

crop$block <- as.factor(crop$block)

crop$fertilizer <- as.factor(crop$fertilizer)

#Applying ANOVA

anova <- aov(yield ~ fertilizer+density+fertilizer:density , data = crop)

anova\_summary<-summary(anova)

anova\_summary