Git Url : <https://github.com/ShilpaAChavan/Datascience-With-R/tree/master/Assignment/Hypothesis%20Testing>

**#################################HypothesisTesting############################**

**#Problem Statement:A F&B manager wants to determine whether there is any significant**

**#difference in the diameter of the cutlet between two units. A randomly selected**

**#sample of cutlets was collected from both units and measured? Analyze the data**

**#and draw inferences at 5% significance level. Please state the assumptions and**

**#tests that you carried out to check validity of the assumptions.**

**# Data : Cutlets.csv**

**##############################################################################**

CutletData <- read.csv(file.choose()) #Cutlets.csv

View(CutletData)

attach(CutletData)

#The attach function allows to access variables of a data.frame without calling the data.frame

#CutletData$Unit.A, Unit.A

#x: cutlet - Discrete datatype

#Y: Unit.A, Unit.B diameter - Continous datatype

# Compare the two population Unit.A and Unit.B

################# Normality Test ###############################################

#Ho : Data is normally distributed

#Ha : Data is not normally distributed.

shapiro.test(Unit.A)

#p-value =0.32 > 0.05 So p high null fly => It follows Normal Distribution.

shapiro.test(Unit.B)

#p-value = 0.52 > 0.05 So p high null fly => It follows Normal Distribution.

################### Variance Test ##########################################

#Ho : var(Unit.A) = var(Unit.B)

#Ha : var(Unit.A) != var(Unit.B)

var.test(Unit.A,Unit.B)

#p-value = 0.31 > 0.05 So p high null fly accept null hypthesis

# Hence variance of Unit.A equal to variance Unit.B

################# 2- Sample T Test #########################################

#Ho : mean(Unit.A) = mean(Unit.B) i.e equal means

#Ha : mean(Unit.A) != mean(Unit.B) i.e not equal means

# Signifiance level = 5% i.e alpha =0.05

?t.test

t.test(Unit.A,Unit.B,alternative = "two.sided",conf.level = 0.95,correct = True)

# alternative = "two.sided" means we are checking for equal and unequal means

#p-value = 0.472 > 0.05 So p high null fly => equal means

#**Conclusion:** Accept Null hypothesis There is no significant difference in the diameter

# of cutlets of Unit A and Unit B i.e the diameters are same.

**#############################Hypothesis Testing#####################################**

**#Problem Statement:A hospital wants to determine whether there is any difference**

**#in the average Turn Around Time (TAT) of reports of the laboratories on their**

**#preferred list. They collected a random sample and recorded TAT for reports**

**#of 4 laboratories. TAT is defined as sample collected to report dispatch.**

**#Analyze the data and determine whether there is any difference in average TAT**

**#among the different laboratories at 5% significance level.**

**# Data : LabTAT.csv**

**#################################################################################**

LabTATData <- read.csv(file.choose()) #LabTAT.csv

View(LabTATData)

colnames(LabTATData)<-c("Lab1","Lab2","Lab3","Lab4")#Change column names

View(LabTATData)

attach(LabTATData)

#The attach function allows to access variables of a data.frame without calling the data.frame

#LabTATData$Laboratory.1, Lab1

#X: lab reports i.e. discrete

#Y: Difference in avg TAT 0f 4 labs Continuous

################# Normality Test ###############################################

#Ho : Data is normally distributed

#Ha : Data is not normally distributed.

shapiro.test(Lab1)

#p-value = 0.55 > 0.05 So p high null fly =>Accept Ho

shapiro.test(Lab2)

#p-value = 0.86 > 0.05 So p high null fly => Accept Ho

shapiro.test(Lab3)

#p-value = 0.42 > 0.05 So p high null fly => Accept Ho

shapiro.test(Lab4)

#p-value = 0.66 > 0.05 So p high null fly => Accept Ho i.e Data is normally distributed

################### Variance Test ##########################################

#Ho: Variance for Lab1,Lab2,Lab3,Lab4 are equal.

#Ha: Variance for Lab1,Lab2,Lab3,Lab4 are not qual.

#bartlett.test for checking variance of more than 2 population.

Stacked\_Data <- stack(LabTATData)

View(Stacked\_Data)

attach(Stacked\_Data)

?bartlett.test

?aov

bartlett.test(values~ind,data=Stacked\_Data) #test for homogeneity of variances

#p-value = 0.106 > 0.05 p high null fly => accept null hypothesis.

#Hence Variances of labs are equal.

#As there are more than 2 Discrete variables and output variable y is continuous

#We carry One way ANOVA Test

#Ho: Mean TAT for all labs are same.

#Ha: Mean TAT for all labs is not same.

View(Stacked\_Data)

anova\_result <- aov(values~ind,data = Stacked\_Data)

summary(anova\_result)

# p-value= 2e-16 = 0 < 0.05 Reject Null hypothesis Mean TAT for all labs are

# not same.

# **Conclusion:**Hence there is difference in average TAT among different laboratories.

**##############################HypothesisTesting#####################################**

**#Problem Statement:Sales of products in four different regions is tabulated for males and females.**

**#Find if male-female buyer rations are similar across regions.**

**# Data : BuyerRatio.csv**

**##################################################################################**

BuyerRatio <- read.csv(file.choose()) #BuyerRatio.csv

attach(BuyerRatio)

View(BuyerRatio)

region1 <- East

region2 <- West

region3 <- North

region4 <- South

region= data.frame(region1,region2,region3,region4)

View(region)

#Ho : Proportions of male and female are same.

#Ha : Proportions of male and female are not same.

?chisq.test

chisq.test(region)

# p-value = 0.66 > 0.05 p high null fly => accept null hypothesis

# Hence the male female buyer ratio is similar across region.

##Second Approach:

stacked\_data<- stack(BuyerRatio)

View(stacked\_data)

attach(stacked\_data)

chisq.test(table(values,ind))

#p-value = 0.2931 > 0.05 p high null fly =>accept null hypothesis

#Hence the male female buyer ratio is similar across region.

#Warning message:

# In chisq.test(table(values, ind)) :

# Chi-squared approximation may be incorrect

#simulate.p.value = TRUE

#data: table(values, ind)

#X-squared = 24, df = NA, p-value = 1

# **Conclusion:** Hence the male female buyer ratio is similar across region.

**###############################Hypothesis Testing################################**

**#Problem Statement:TeleCall uses 4 centers around the globe to process customer order**

**#forms. They audit a certain % of the customer order forms. Any error in order form**

**#renders it defective and has to be reworked before processing. The manager wants to**

**#check whether the defective % varies by centre. Please analyze the data at 5%**

**# significance level and help the manager draw appropriate inferences**

**# Data : Costomer+OrderForm.csv**

**##################################################################################**

library(readxl)

customerData <- read.csv(file.choose()) #Costomer+OrderForm.csv

View(customerData)

customOrder <- as.data.frame(lapply(customerData, as.numeric))

View(customOrder)

#Ho : Defective percentage are same across center.

#Ha : Defective percentage are varies across center.

?chisq.test

chisq.test(customOrder)

#data: customOrder

#X-squared = 44.655, df = 897, p-value = 1

#Warning message:

#In chisq.test(customOrder) : Chi-squared approximation may be incorrect

chisq.test(customOrder,simulate.p.value = TRUE)

#data: customOrder

#X-squared = 44.655, df = NA, p-value = 1

#p-value = 1 >0.05 so p high null fly => accept NULL hypothesis

#**Conclusion:** Hence the defective percentage are same across the centers.

**QUERY: Didn’t got this simulate p-value concept. Also why the df= NA for chisq.test with simulate.p.value = TRUE else it is 897. Please reply.**

**#############################Hypothesis Testing#####################################**

**#Problem Statement:Fantaloons Sales managers commented that % of males versus females**

**#walking in to the store differ based on day of the week. Analyze the data and determine**

**#whether there is evidence at 5 % significance level to support this hypothesis.**

**# Data : Faltoons.csv**

**#################################################################################**

faltoonData <- read.csv(file.choose()) #Faltoons.csv

attach(faltoonData)

View(faltoonData)

table1 <- table(Weekdays,Weekend)

View(table1)

table1

weekdays <- (table(Weekdays))

weekdays

#Female Male

#287 113

weekend <- table(Weekend)

weekend

#Female Male

#233 167

# Both input and output variables are discrete. We have 2 population to compare so we

#procced with 2- proportion test.

#Ho : Proportions of male and female in weekdays are same.

#Ha : Proportions of male and female in weekdays are different.

prop.test(x=c(287,167),n=c(520,270),conf.level = 0.95,correct = FALSE,alternative = "two.sided")

# two. sided -> means checking for equal proportions and unequal proportions of Male and Femal for weekdays

# p-value = 0.07254 > 0.05 so p high null fly => accept null hypothesis.

#Hence proportions of males and female on weekdays are same.

#Ho : Proportions of male and female in weekend are same.

#Ha : Proportions of male and female in weekend are different.

prop.test(x=c(233,113),n=c(520,270),conf.level = 0.95,correct = FALSE,alternative = "two.sided")

# two. sided -> means checking for equal proportions and unequal proportions of Male and Femal for weekdays

# p-value = 0.4271 > 0.05 so p high null fly => accept null hypothesis.

#Hence proportions of males and female on weekend are same.

#Hence we conclude that the proportion of males and females are same on weekdays and weekend respectively.