**Define :-**

A magazine does a survey and wants to publish results. The following data represent business startup costs (thousands of dollars) for shops, gathered thru a survey. One of the Questions to be answered is that is it equally expensive to start any of the businesses explored below or are some businesses lighter on the pocket?

Description of variables:-

* X1 = startup costs for pizza
* X2 = startup costs for baker/donuts
* X3 = startup costs for shoe stores
* X4 = startup costs for gift shops
* X5 = startup costs for pet stores

What is the conclusion that the article should publish?

# D – Define the business Y – check if the distribution of startup costs are similar across lines of business

# Import the data

startupcost <- read.csv("H:/springer book/Case study/CaseStudy5/startupcost.csv", stringsAsFactors=FALSE)

View(startupcost)

# C and O – not required for this project

# V – Visualise the data to establish if the data is parametric or non parametric . This sill influence the tests .

str(startupcost)

> str(startupcost)

'data.frame': 38 obs. of 5 variables:

$ X1: int 80 125 35 58 110 140 97 50 65 79 ...

$ X2: int 150 40 120 75 160 60 45 100 86 87 ...

$ X3: int 48 35 95 45 75 115 42 78 65 125 ...

$ X4: int 100 96 35 99 75 150 45 100 120 50 ...

$ X5: int 25 80 30 35 30 28 20 75 48 20 ...

startup2 <- startupcost

recon1 <- rowSums(!is.na(startup2[-(1:5)]))

Note :- No NA values show up in the dataframe

# check for EQUAL variance

> var.test(startup2$X1,startup2$X2)

F test to compare two variances

data: startup2$X1 and startup2$X2

F = 0.8088, num df = 12, denom df = 21, p-value = 0.721

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.306730 2.472794

sample estimates:

ratio of variances

0.8087739

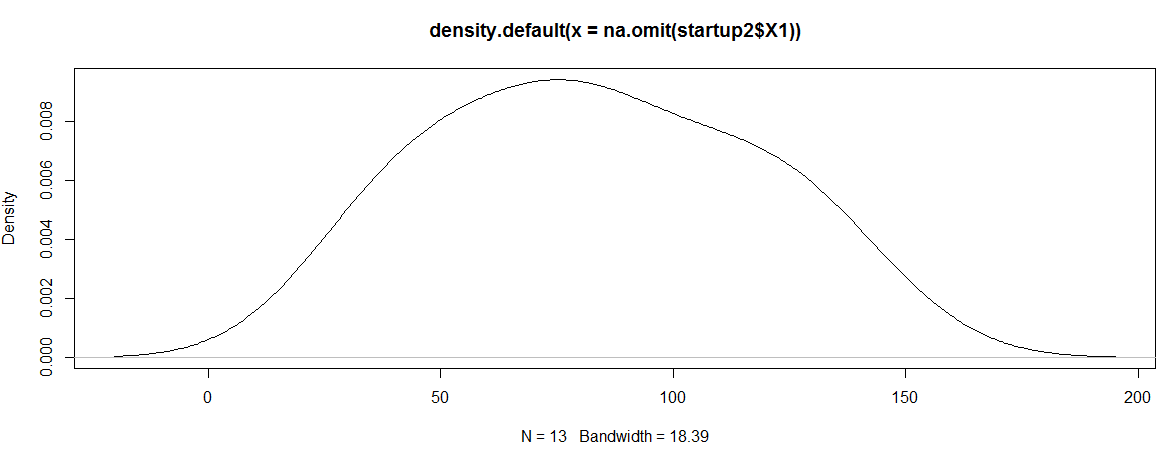
Note :- We cannot REJECT the null because p value >.05. Hence we have to ACCEPT the null that variance is equal .

**Problem for you :- Please do this for the other variables in the dataframe and conclude on the variance**

# Normal distribution

d <- density(na.omit(startup2$X1))

plot(d)



library("pastecs")

stat.desc(na.omit(startup2$X1))

> stat.desc(na.omit(startup2$X1))

x

nbr.val 13.0000000

nbr.null 0.0000000

nbr.na 0.0000000

min 35.0000000

max 140.0000000

range 105.0000000

sum 1079.0000000

median 80.0000000

mean 83.0000000

SE.mean 9.4672174

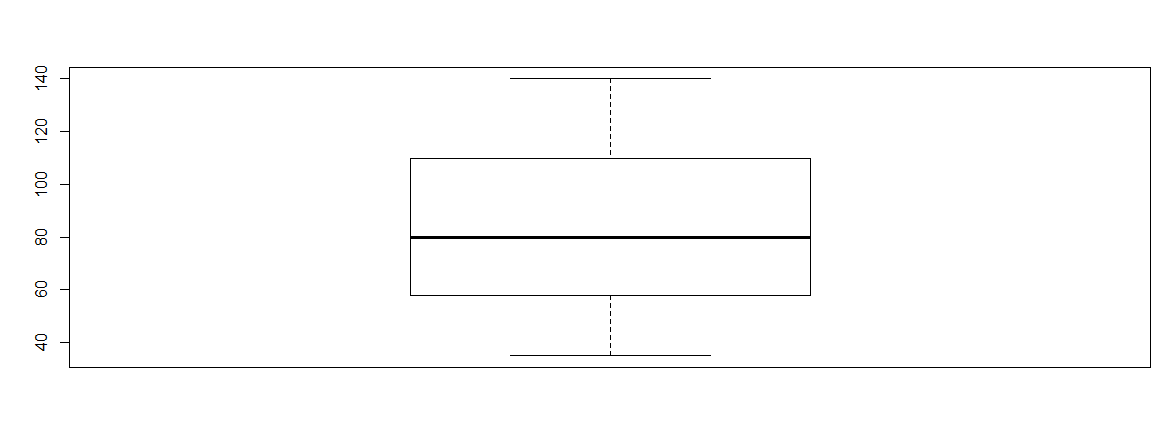
CI.mean.0.95 20.6272947

var 1165.1666667

std.dev 34.1345377

coef.var 0.4112595

boxplot(startup2$X1)



Conclusion :- Mean is ~ Median and the boxplot shows a normal distribution

**Problem for you :- Please do this for the other variables in the dataframe and conclude on the distribution**

# A – Analyse . Run the t tests and ANOVA and Chi sq test

#T tests - Equal Variance

# t test

t.test(startup2$X1 , startup2$X2 ,var.equal = TRUE)

> t.test(startup2$X1 , startup2$X2 ,var.equal = TRUE)

Two Sample t-test

data: startup2$X1 and startup2$X2

t = -0.7098, df = 33, p-value = 0.4828

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-35.14897 16.96715

sample estimates:

mean of x mean of y

83.00000 92.09091

Conclusion :- If the p value is less than equal to .05 then we can REJECT the null. Here we will have to ACCEPT the null that the means are equal

#T tests – Un-equal Variance

t.test(startup2$X1 , startup2$X2 ,var.equal = FALSE)

**Problem for you :- On the basis of the outcome of the test for equal variance, run the T tests – Un-equal Variance**

# ANOVA

str(startup2)

data1<- aov(X1 ~ X2+X3+X4, data=startup2)

summary(data1)

> summary(data1)

Df Sum Sq Mean Sq F value Pr(>F)

X2 1 1572 1571.6 1.196 0.303

X3 1 180 180.5 0.137 0.719

X4 1 405 404.7 0.308 0.592

Residuals 9 11825 1313.9

25 observations deleted due to missingness

Conclusion: - If the p value is less than equal to .05 then we can REJECT the null. Here we will have to ACCEPT the null that the means are equal

Assumptions:-

1. Residuals should be normally distributed

data1$residuals

> data1$residuals

1 2 3 4 5 6

13.281351 31.464608 -36.297836 -28.239742 45.878150 29.793708

7 8 9 10 11 12

12.354359 -34.993888 -25.068705 -8.550637 -47.857625 21.052730

13

27.183529

shapiro.test(data1$residuals)

> shapiro.test(data1$residuals)

Shapiro-Wilk normality test

data: data1$residuals

W = 0.9164, p-value = 0.2242

Conclusion :- Since p value is greater than .05 we cannot reject the null and have to ACCEPT the null that the data is normally distributed

1. Variances are equal across the 4 samples considered

bartlett.test(list(startup2$X1,startup2$X2,startup2$X3,startup2$X4))

> bartlett.test(list(startup2$X1,startup2$X2,startup2$X3,startup2$X4))

Bartlett test of homogeneity of variances

data: list(startup2$X1, startup2$X2, startup2$X3, startup2$X4)

Bartlett's K-squared = 0.9174, df = 3, p-value = 0.8212

Conclusion :- Since p value is greater than .05 we cannot reject the null and have to ACCEPT the null that the data variables have equal vaariances

**Bartlett’s test**

Bartlett’s test allows you to compare the variance of two or more samples to determine whether they are drawn from populations with equal variance. It is suitable for normally distributed data. The test has the null hypothesis that the variances are equal and the alternative hypothesis that they are not equal.

This test is useful for checking the assumptions of an analysis of variance.

**Problem for you: - Practice ANOVA considering all 5 variables**

# Insight generation

> summary(startup2)

X1 X2 X3 X4

Min. : 35 Min. : 40.00 Min. : 35.0 Min. : 35.0

1st Qu.: 58 1st Qu.: 63.75 1st Qu.: 45.0 1st Qu.: 50.0

Median : 80 Median : 87.00 Median : 70.0 Median : 97.5

Mean : 83 Mean : 92.09 Mean : 72.3 Mean : 87.0

3rd Qu.:110 3rd Qu.:115.00 3rd Qu.: 95.0 3rd Qu.:100.0

Max. :140 Max. :160.00 Max. :125.0 Max. :150.0

NA's :25 NA's :16 NA's :18 NA's :18

X5

Min. : 20.00

1st Qu.: 29.50

Median : 49.00

Mean : 51.62

3rd Qu.: 75.00

Max. :110.00

The cost of starting

* X1 = startup costs for pizza
* X2 = startup costs for baker/donuts
* X3 = startup costs for shoe stores
* X4 = startup costs for gift shops

Is similar as per result of ANOVA.

**Problem for you: - Practice ANOVA considering all 5 variables and write the Insight .**