SELECTED DESCRIPTIVE AND INFERENTIAL STATISTICS IN R FOR UNDERGRADUATE PROJECT

BEING A PRACTICAL AND DEMONSTRATION SESSSION HELD ON THURSDAY, 18TH MAY, 2023 AT STATISTICS DEPARTMENT, THE FEDERAL POLYTECHNIC OFFA

BY DR. ABDULKABIR MURITALA

# Descriptives Statistics simulation

x=rnorm(100)  
mean(x)

## [1] 0.1227049

var(x)

## [1] 1.054929

sd(x)

## [1] 1.027097

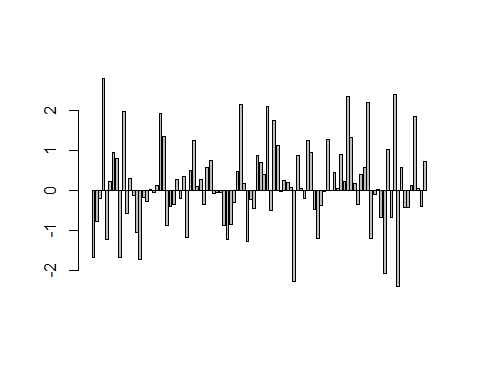
median(x)

## [1] 0.04950565

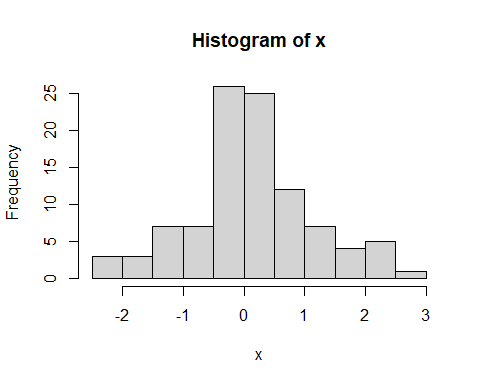
summary(x)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.38564 -0.41825 0.04951 0.12270 0.69701 2.78144

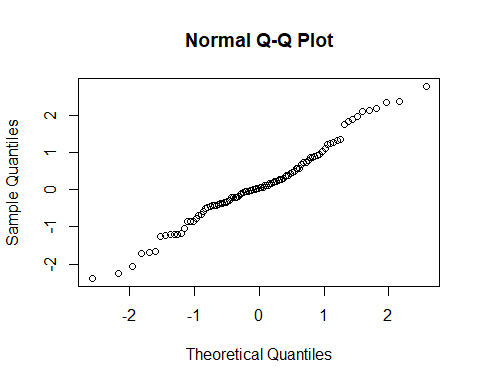
barplot(x)



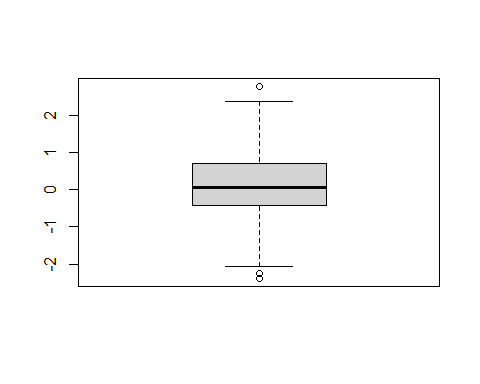
hist(x)



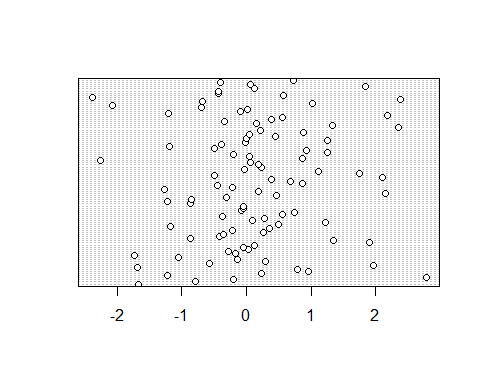
qqnorm(x)



boxplot(x)

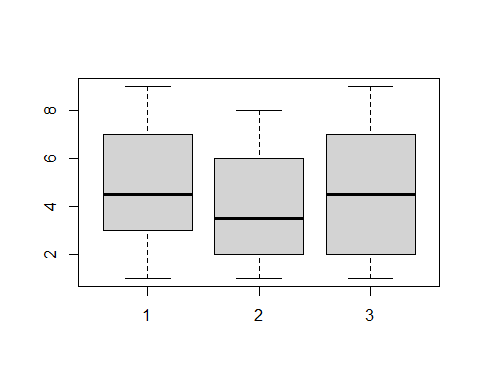


dotchart(x)

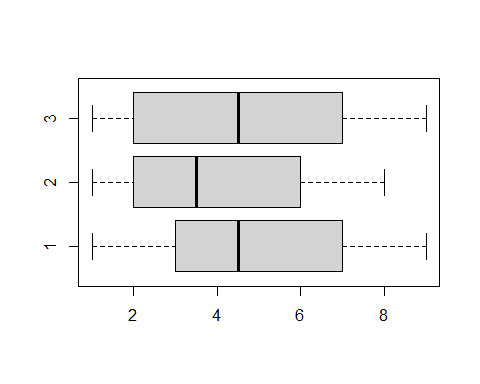


#Real life data

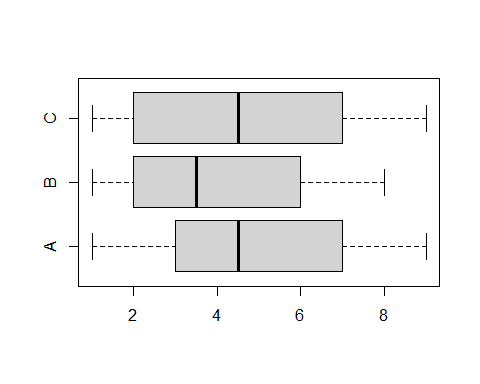
x=c(1,3,5,6,7,8,9,3,2,4)  
y=c(2,3,5,6,7,8,4,3,2,1)  
z=c(2,3,4,1,2,5,6,7,8,9)  
boxplot(x,y,z)



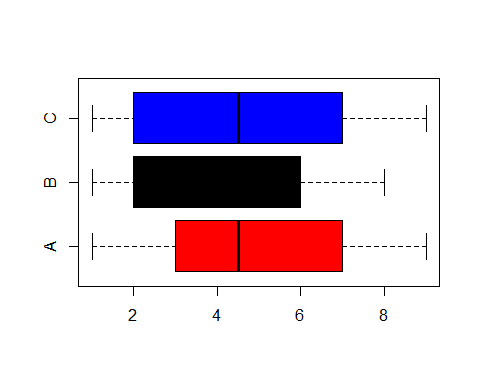
boxplot(x,y,z,horizontal=T)



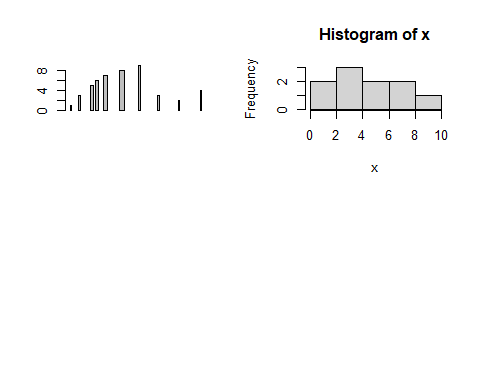
boxplot(x,y,z,horizontal=T,names=c("A","B","C"))



boxplot(x,y,z,horizontal=T,names=c("A","B","C"),col=c("red","black","blue"))



par(mfrow=c(2,2))  
barplot(x,y,z)  
hist(x)



#Inferential Statistics ##chi sqaure

data <- matrix(c(100, 70, 20, 90, 75, 25), ncol=3, byrow=TRUE)  
colnames(data) <- c("Rep","Dem","Ind")  
rownames(data) <- c("Male","Female")  
data <- as.table(data)  
data

## Rep Dem Ind  
## Male 100 70 20  
## Female 90 75 25

chisq.test(data)

##   
## Pearson's Chi-squared test  
##   
## data: data  
## X-squared = 1.2543, df = 2, p-value = 0.5341

#one sample t.test

daily.sales=c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)

mean(daily.sales)

## [1] 6753.636

sd(daily.sales)

## [1] 1142.123

quantile(daily.sales)

## 0% 25% 50% 75% 100%   
## 5260 5910 6515 7515 8770

t.test(daily.sales)

##   
## One Sample t-test  
##   
## data: daily.sales  
## t = 19.612, df = 10, p-value = 2.599e-09  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 5986.348 7520.925  
## sample estimates:  
## mean of x   
## 6753.636

t.test(daily.sales,mu=7000)

##   
## One Sample t-test  
##   
## data: daily.sales  
## t = -0.71542, df = 10, p-value = 0.4907  
## alternative hypothesis: true mean is not equal to 7000  
## 95 percent confidence interval:  
## 5986.348 7520.925  
## sample estimates:  
## mean of x   
## 6753.636

#Two sample

dat=read.csv("hnd.csv")  
attach(dat)  
t.test(exp~sex,var.equal=T)

##   
## Two Sample t-test  
##   
## data: exp by sex  
## t = -0.97163, df = 9, p-value = 0.3566  
## alternative hypothesis: true difference in means between group F and group M is not equal to 0  
## 95 percent confidence interval:  
## -3.328219 1.328219  
## sample estimates:  
## mean in group F mean in group M   
## 5 6

#Paired sample t-test

dat=read.csv("hnd2.csv")  
t.test(dat$pre,dat$post,paired=T)

##   
## Paired t-test  
##   
## data: dat$pre and dat$post  
## t = 4.3846, df = 9, p-value = 0.001759  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## 0.9197319 2.8802681  
## sample estimates:  
## mean difference   
## 1.9

#Regression

x1=runif(50,2,3)  
x2=rnorm(50,4,1)  
y=2\*x1+x2  
dat=data.frame(y,x1,x2)  
m1=lm(y~x1+x2,dat)  
m1

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = dat)  
##   
## Coefficients:  
## (Intercept) x1 x2   
## -8.039e-15 2.000e+00 1.000e+00

summary(m1)

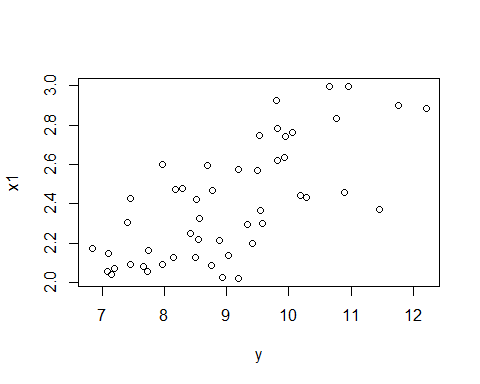
## Warning in summary.lm(m1): essentially perfect fit: summary may be unreliable

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.229e-14 -5.870e-17 2.909e-16 7.206e-16 1.507e-15   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.039e-15 2.322e-15 -3.463e+00 0.00115 \*\*   
## x1 2.000e+00 1.004e-15 1.991e+15 < 2e-16 \*\*\*  
## x2 1.000e+00 2.976e-16 3.360e+15 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.922e-15 on 47 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.124e+31 on 2 and 47 DF, p-value: < 2.2e-16

predict(m1)

## 1 2 3 4 5 6 7 8   
## 7.726476 8.498442 9.545283 9.340590 10.887471 8.143632 11.756041 7.658687   
## 9 10 11 12 13 14 15 16   
## 8.538179 8.520185 8.567244 9.501465 9.820067 7.400834 9.407096 8.774894   
## 17 18 19 20 21 22 23 24   
## 9.194433 8.176168 10.190122 8.756397 9.520905 10.058001 11.449866 6.834045   
## 25 26 27 28 29 30 31 32   
## 8.414048 10.761051 7.140366 9.194395 8.294391 9.805767 9.820311 7.458458   
## 33 34 35 36 37 38 39 40   
## 9.580948 10.652683 9.028489 8.889347 7.076613 7.970347 10.278749 10.964060   
## 41 42 43 44 45 46 47 48   
## 8.924511 9.928700 7.097705 7.196671 12.209903 7.741069 8.696273 9.942406   
## 49 50   
## 7.457211 7.969935

plot(y,x1)

 #Correlation

x1=rnorm(100,2,1)  
x2=rnorm(100,2,4)  
cor(x1,x2)

## [1] -0.0840994

cor.test(x1,x2)

##   
## Pearson's product-moment correlation  
##   
## data: x1 and x2  
## t = -0.8355, df = 98, p-value = 0.4055  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2759590 0.1142052  
## sample estimates:  
## cor   
## -0.0840994

cor.test(x1,x2,method="spearman")

##   
## Spearman's rank correlation rho  
##   
## data: x1 and x2  
## S = 182520, p-value = 0.3454  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.09522952

cor.test(x1,x2,method="kendall")

##   
## Kendall's rank correlation tau  
##   
## data: x1 and x2  
## z = -1.0304, p-value = 0.3028  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
## tau   
## -0.06989899

#One way ANOVA

x1=runif(50,2,3)  
y=2\*x1+x1  
anova(lm(y~x1))

## Warning in anova.lm(lm(y ~ x1)): ANOVA F-tests on an essentially perfect fit  
## are unreliable

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 36.719 36.719 1.2329e+32 < 2.2e-16 \*\*\*  
## Residuals 48 0.000 0.000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Two way Anova

x1=runif(50,2,3)  
x2=rnorm(50,4,1)  
y=2\*x1+x2  
dat=data.frame(y,x1,x2)  
anova(lm(y~x1+x2,dat))

## Warning in anova.lm(lm(y ~ x1 + x2, dat)): ANOVA F-tests on an essentially  
## perfect fit are unreliable

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 16.162 16.162 3.7701e+31 < 2.2e-16 \*\*\*  
## x2 1 57.899 57.899 1.3506e+32 < 2.2e-16 \*\*\*  
## Residuals 47 0.000 0.000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Two way Anova with replication

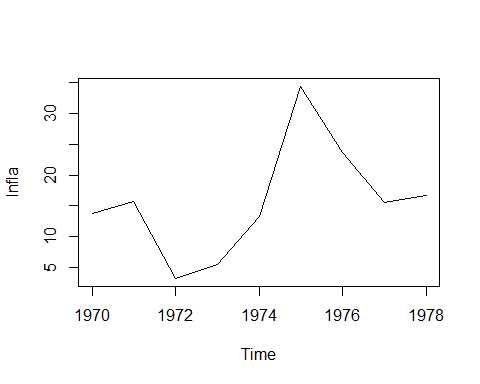
x1=runif(50,2,3)  
x2=rnorm(50,4,1)  
y=2\*x1+x2  
dat=data.frame(y,x1,x2)  
anova(lm(y~x1\*x2,dat))

## Warning in anova.lm(lm(y ~ x1 \* x2, dat)): ANOVA F-tests on an essentially  
## perfect fit are unreliable

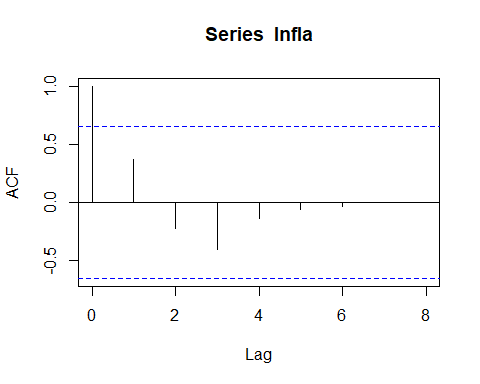
## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 10.844 10.844 7.7001e+29 <2e-16 \*\*\*  
## x2 1 37.825 37.825 2.6860e+30 <2e-16 \*\*\*  
## x1:x2 1 0.000 0.000 1.7720e-01 0.6758   
## Residuals 46 0.000 0.000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Time series

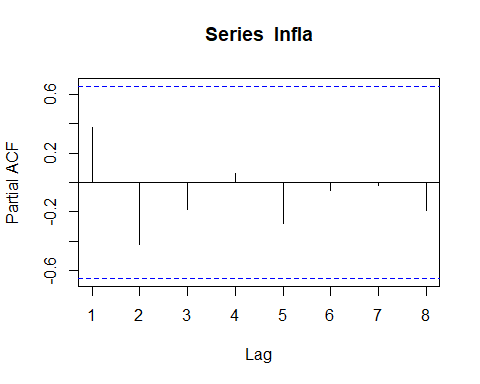
Infla=ts(c(13.8,15.7,3.2,5.4,13.2,34.4,23.7,15.6,16.6),start= c(1970,1))  
plot(Infla)



acf(Infla)



pacf(Infla)



library(tseries)

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

adf.test(Infla)

##   
## Augmented Dickey-Fuller Test  
##   
## data: Infla  
## Dickey-Fuller = -4.3187, Lag order = 2, p-value = 0.01214  
## alternative hypothesis: stationary

adf.test(diff(Infla))

##   
## Augmented Dickey-Fuller Test  
##   
## data: diff(Infla)  
## Dickey-Fuller = -1.8109, Lag order = 1, p-value = 0.6444  
## alternative hypothesis: stationary