for statistics

https://andyigg.com/spa3-1/

Software

- R: http://cran.r-project.org/

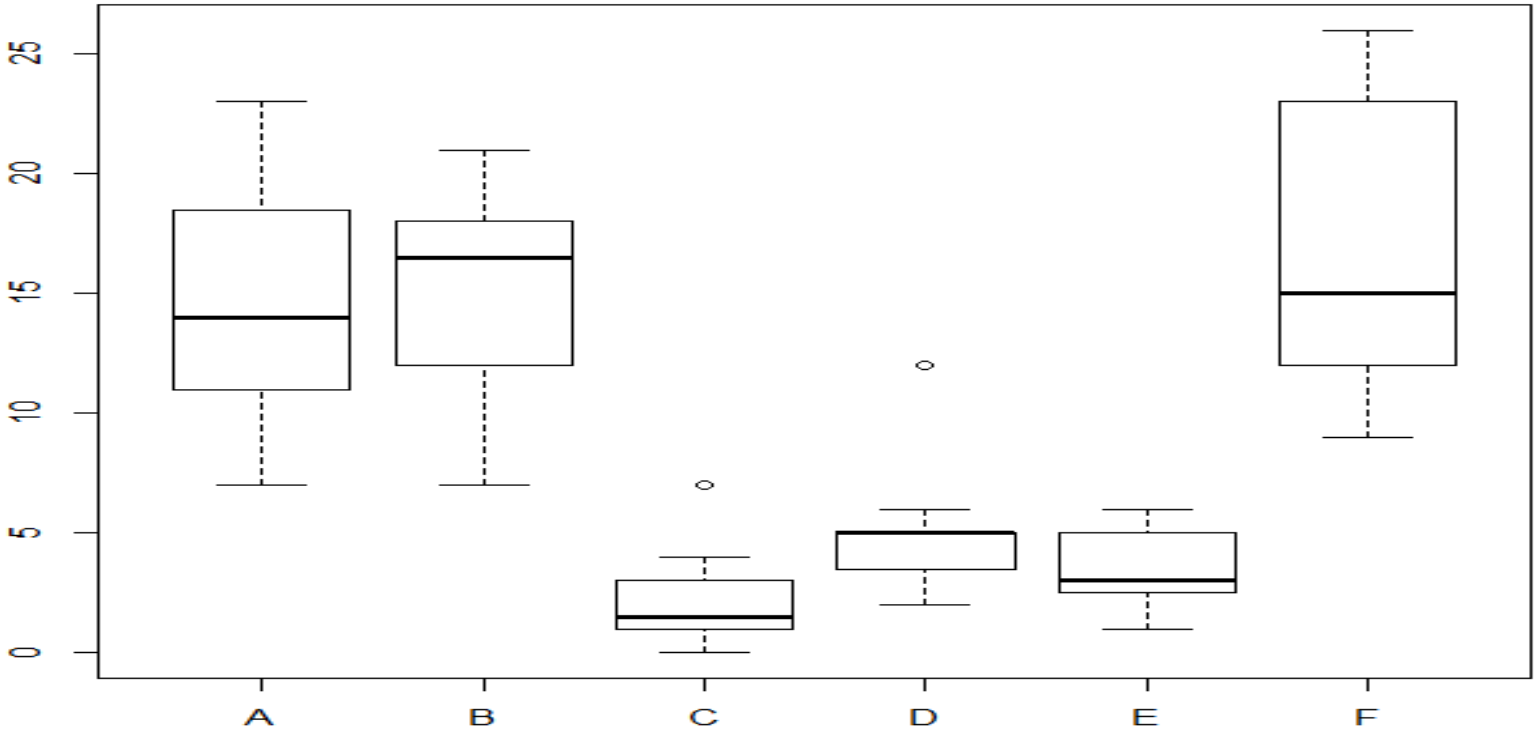
- RStudio: https://www.rstudio.com/products/rstudio/download/#download

Box-and-Whisker Plot

- ?boxplot
- boxplot
- ??InsectSprays

```
> InsectSprays
   count spray
1     10     A
2      7     A
3     20     A
4     14     A
5     14     A
6     12     A
7     10     A
8     23     A
9     17     A
10    20     A
11    14     A
12    13     A
13    11     B
14    17     B
15    21     B
16    11     B
17    16     B
18    14     B
19    17     B
20    17     B
21    19     B
22    21     B
23     7     B
24    13     B
25     0     C
26     1     C
27     7     C
28     2     C
29     3     C
30     1     C
31     2     C
32     1     C
33     3     C
34     0     C
35     1     C
36     4     C
37     3     D
38     5     D
39    12     D
40     6     D
41     4     D
42     3     D
43     5     D
44     5     D
45     5     D
```

boxplot(count ~ spray, data = InsectSprays)



InsectSprays {datasets}

Effectiveness of Insect Sprays

Description

The counts of insects in agricultural experimental units treated with different insecticides.

Usage

```
InsectSprays
```

Format

A data frame with 72 observations on 2 variables.

[,1] count numeric Insect count
[,2] spray factor The type of spray

Source

Beall, G., (1942) The Transformation of data from entomological field experiments, *Biometrika*, **29**, 243–262.

References

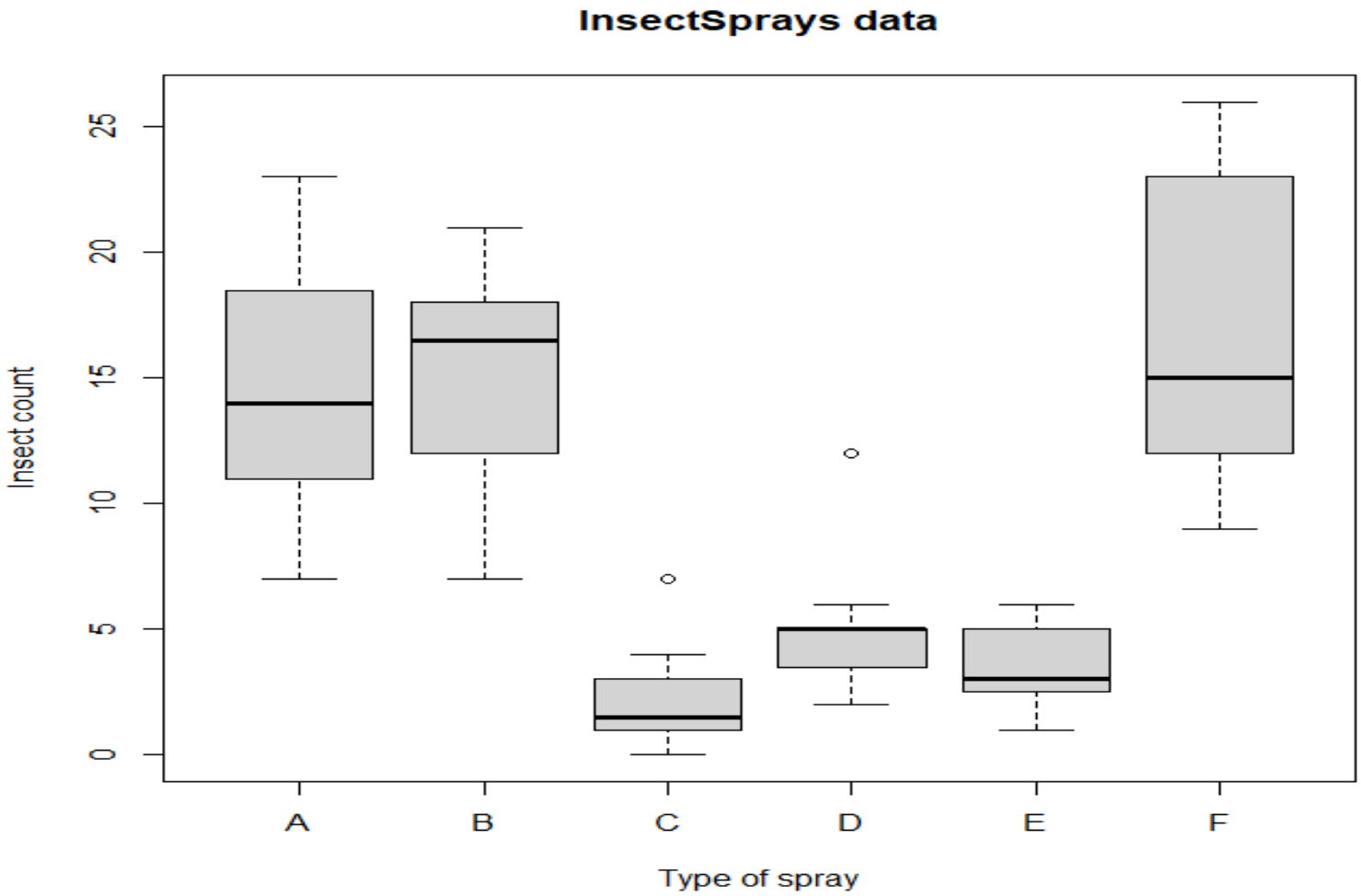
McNeil, D. (1977) *Interactive Data Analysis*. New York: Wiley.

Examples

```
require(stats); require(graphics)
boxplot(count ~ spray, data = InsectSprays,
        xlab = "Type of spray", ylab = "Insect count",
        main = "InsectSprays data", varwidth = TRUE, col = "lightgray")
fm1 <- aov(count ~ spray, data = InsectSprays)
summary(fm1)
opar <- par(mfrow = c(2, 2), oma = c(0, 0, 1.1, 0))
plot(fm1)
fm2 <- aov(sqrt(count) ~ spray, data = InsectSprays)
summary(fm2)
plot(fm2)
par(opar)
```

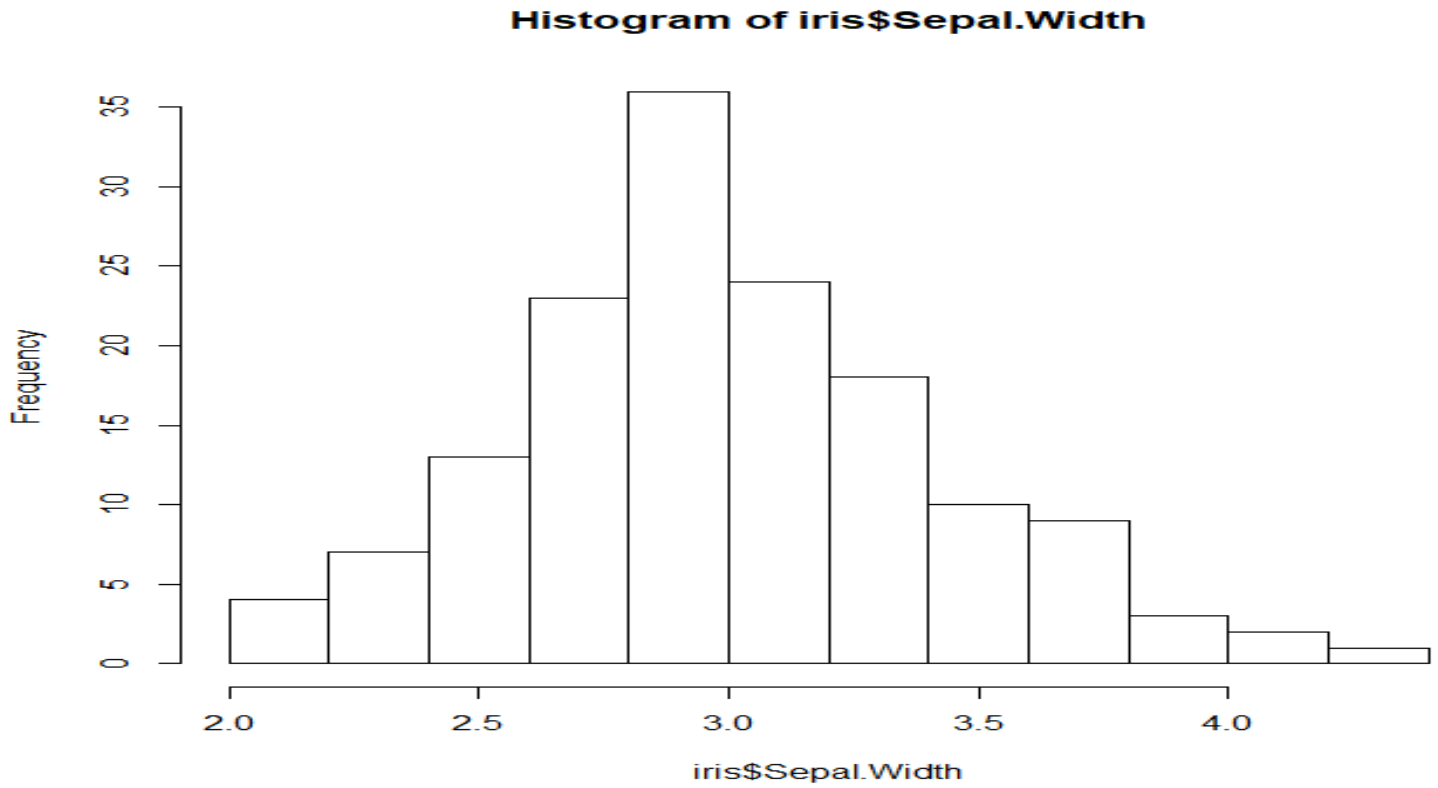
http://127.0.0.1:25229/library/datasets/html/InsectSprays.html

boxplot(count ~ spray, data = InsectSprays, xlab = "Type of spray", ylab = "Insect count", main = "InsectSprays data", varwidth = TRUE, col = "lightgray")

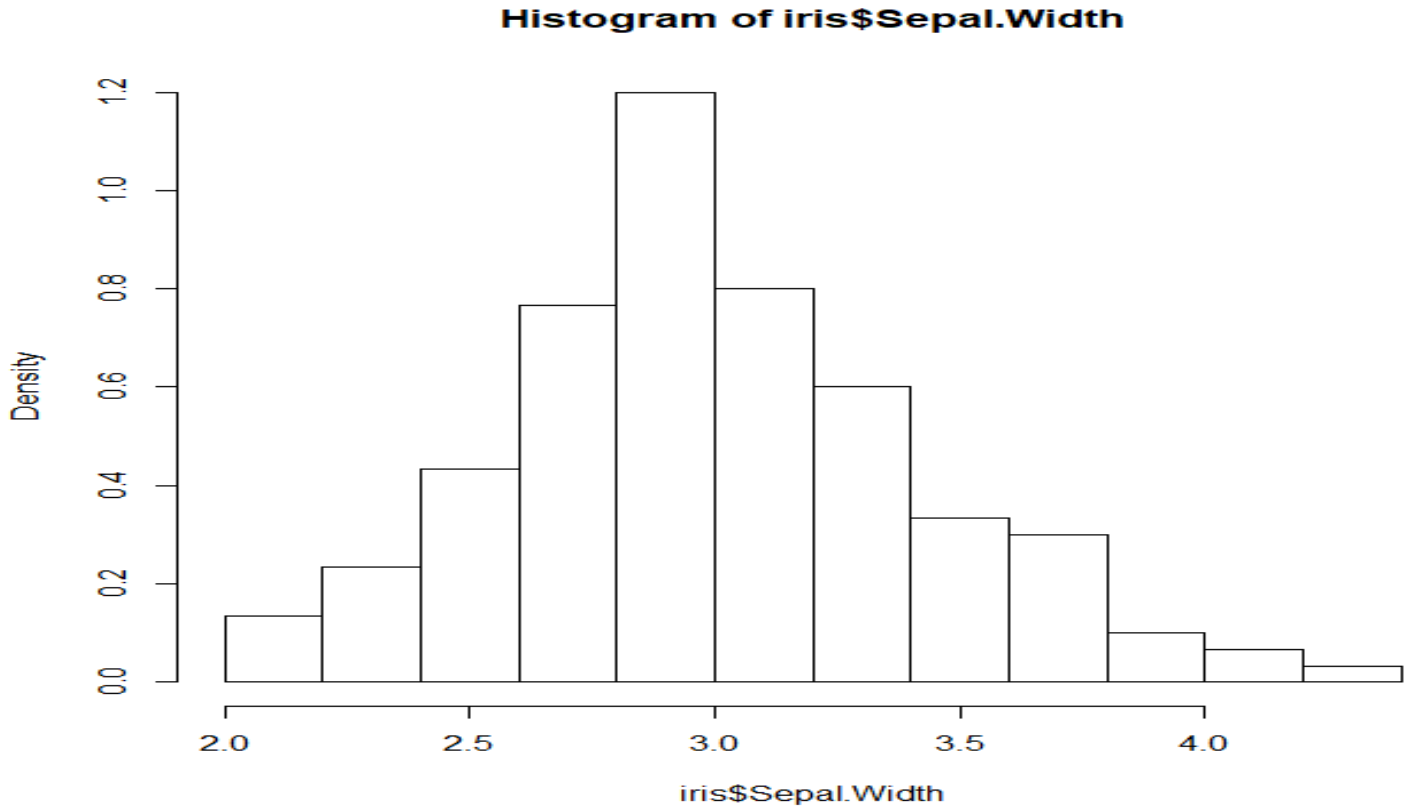


Histograms

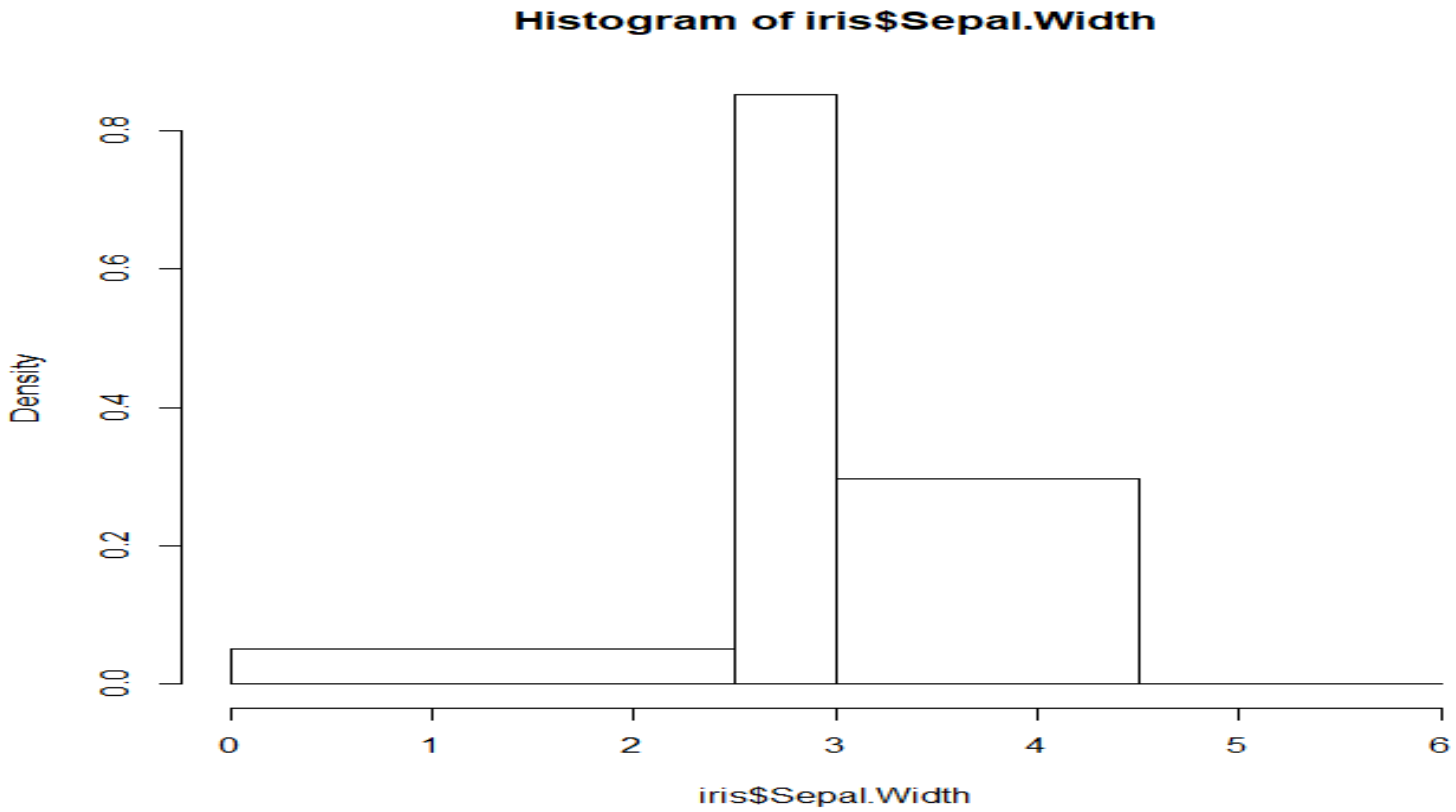
- hist(iris$Sepal.Width)



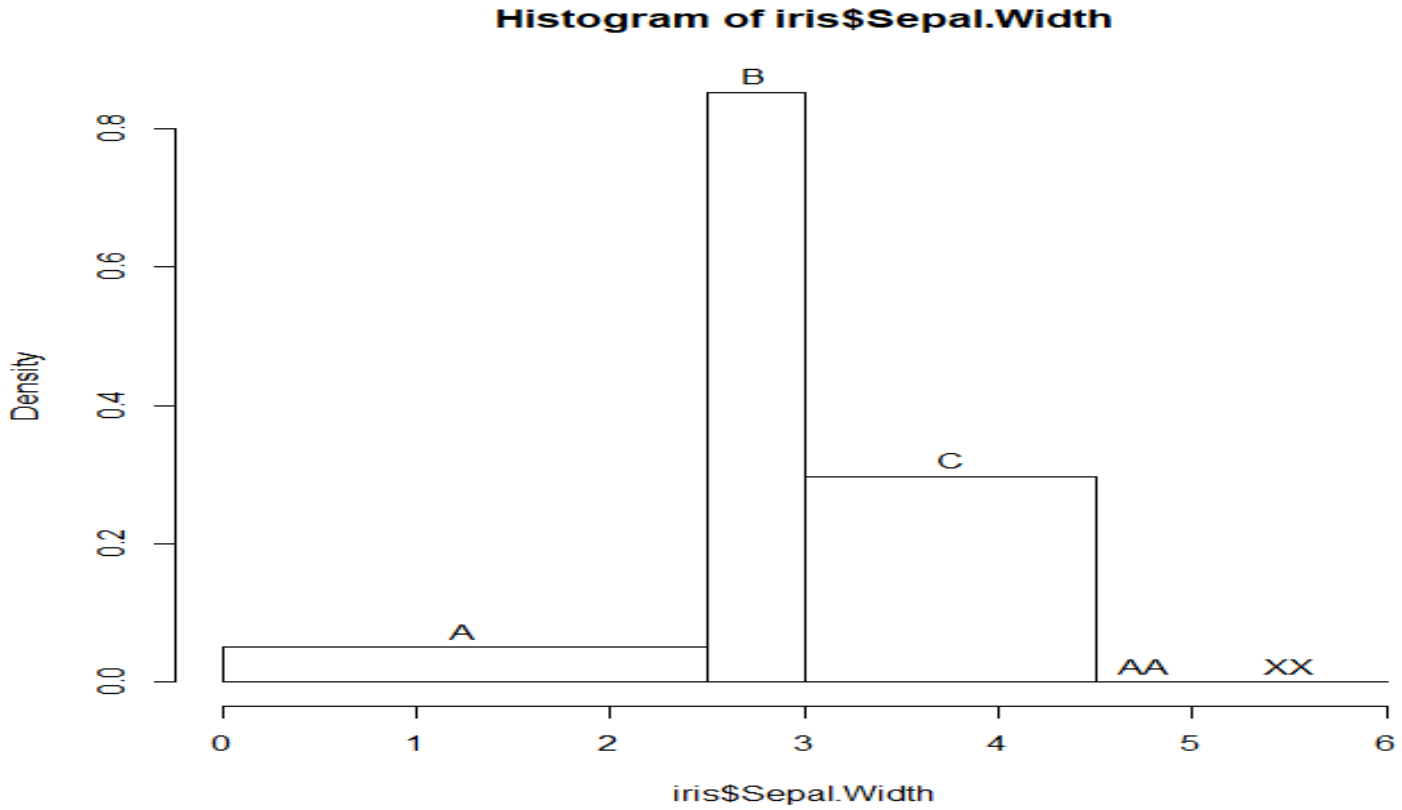
hist(iris$Sepal.Width, freq = FALSE)
sum of probability = 1



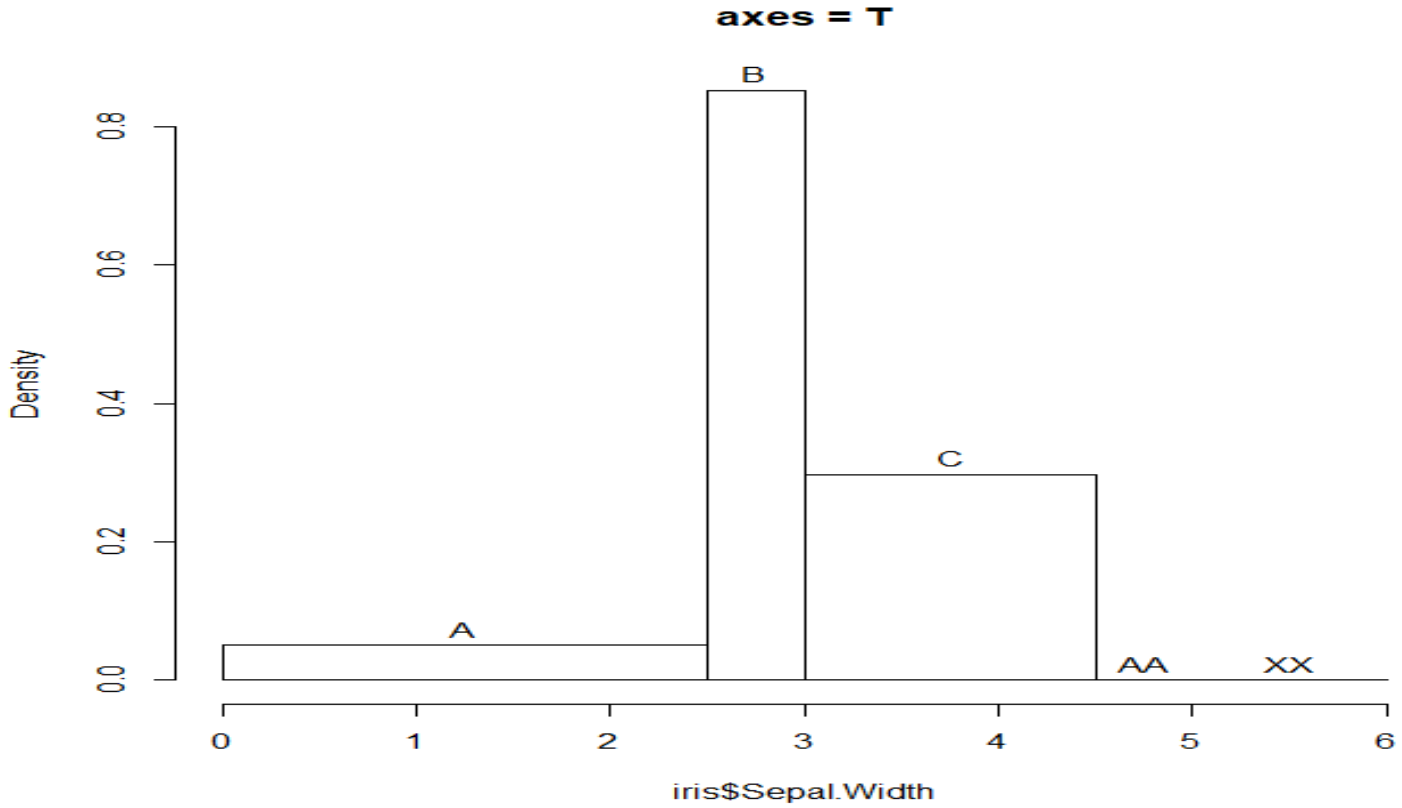
hist(iris$Sepal.Width, freq = FALSE, breaks = c(0, 2.5, 3, 4.5, 5, 6))



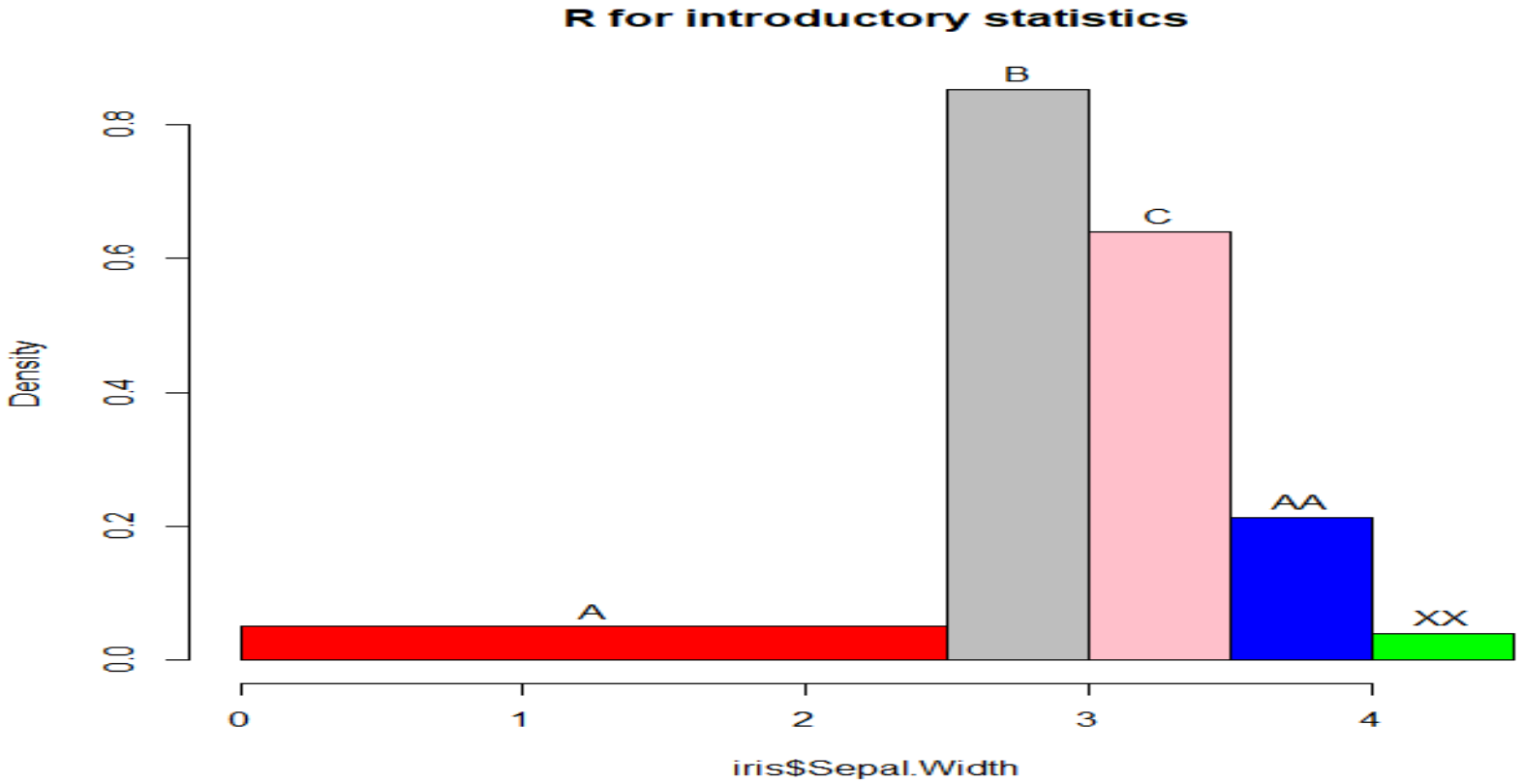
hist(iris$Sepal.Width, freq = FALSE, breaks = c(0, 2.5, 3, 4.5, 5, 6), , labels = c("A", "B", "C", "AA", "XX"))



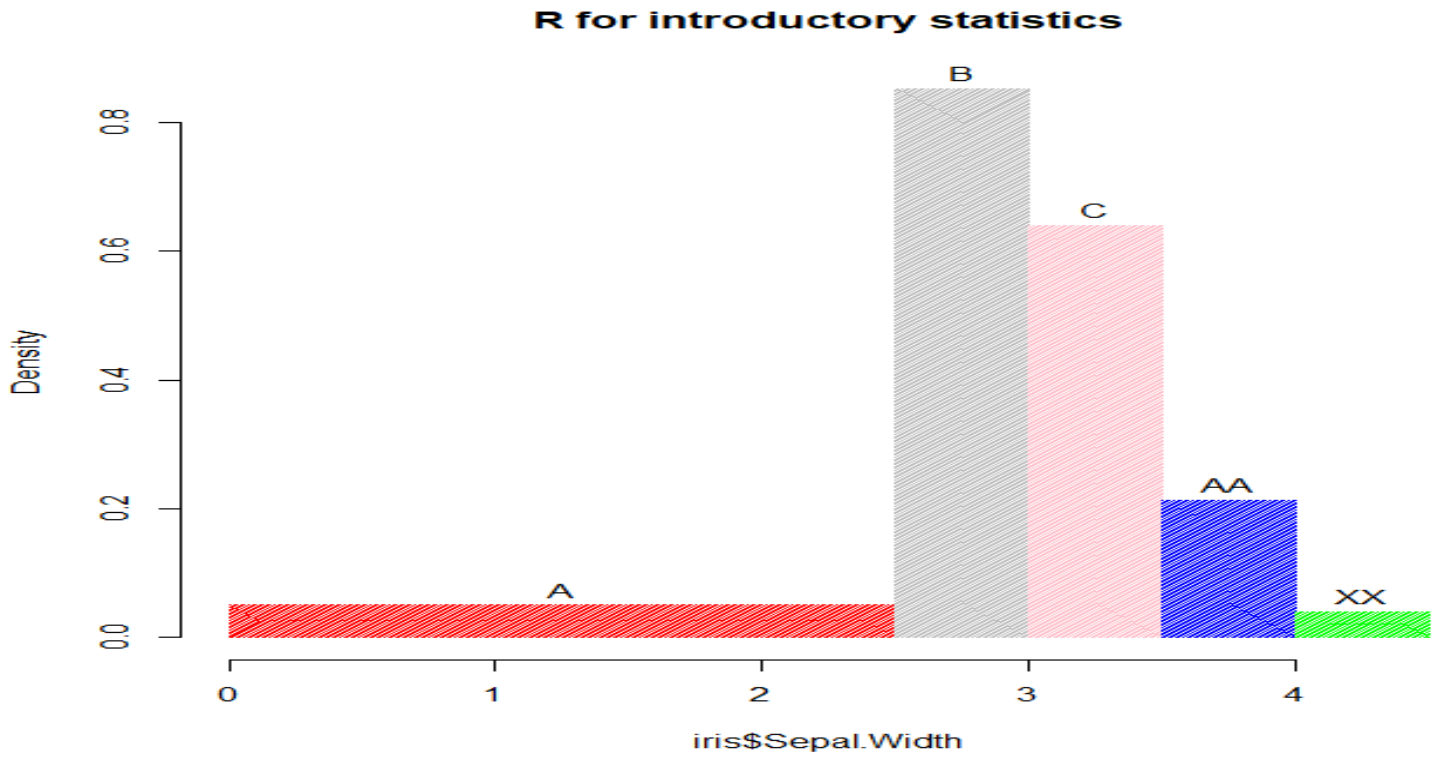
hist(iris$Sepal.Width, freq = FALSE, breaks = c(0, 2.5, 3, 4.5, 5, 6), , labels = c("A", "B", "C", "AA", "XX"), main = "axes = T")



```
hist(iris$Sepal.Width, freq = FALSE, breaks = c(0, 2.5, 3, 3.5, 4, 4.5), ,
                    labels = c("A", "B", "C", "AA", "XX"),
                  main = "R for introductory statistics" , col = c("red", "gray",
                    "pink", "blue", "green")     )
```



```
hist(iris$Sepal.Width, freq = FALSE, breaks = c(0, 2.5, 3, 3.5, 4, 4.5), ,
                    labels = c("A", "B", "C", "AA", "XX"),
                  main = "R for introductory statistics" , col = c("red", "gray",
                                "pink", "blue", "green"),
                   border = NA,            density = 86                          )
```



Stem-and-leaf diagram

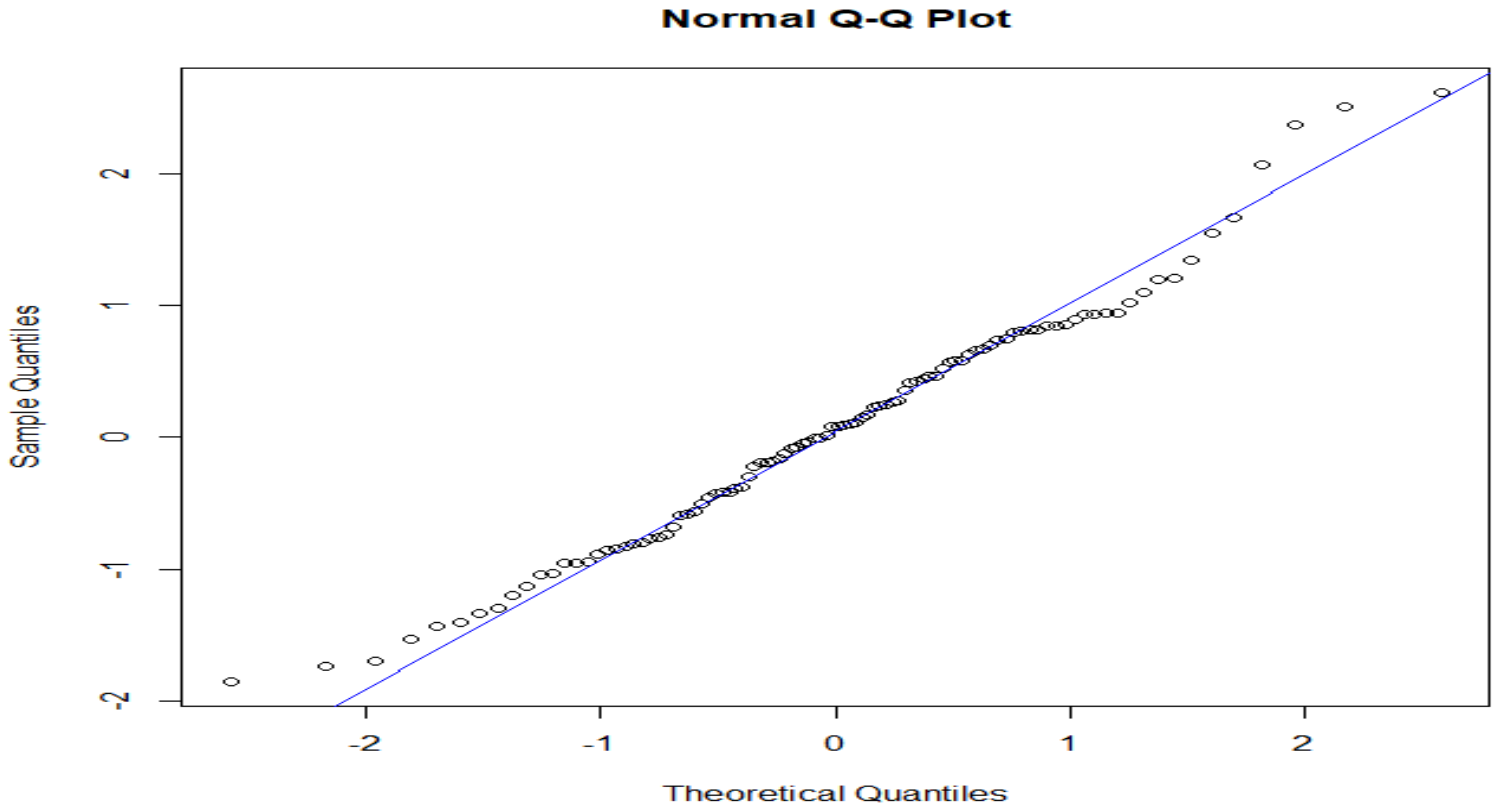
- stem(x,scale=1,width=80,atom=1e-08)
- data(faithful) # Old Faithful geyser at Yellowstone National Park, 272 observations on 2variables, eruptions and waiting

```
> stem(faithful$waiting)

  The decimal point is 1 digit(s) to the right of the |

  4 | 3
  4 | 55566666777788899999
  5 | 00000111111222223333333444444444
  5 | 55555566667778888999999
  6 | 0000002222333444 4
  6 | 555667899
  7 | 00001111123333333444444
  7 | 55555555666666666777777777777888888888888888999999999
  8 | 000000001111111111111222222222223333333333333344444444 4
  8 | 555555666666778888888999
  9 | 00000012334
  9 | 6
```

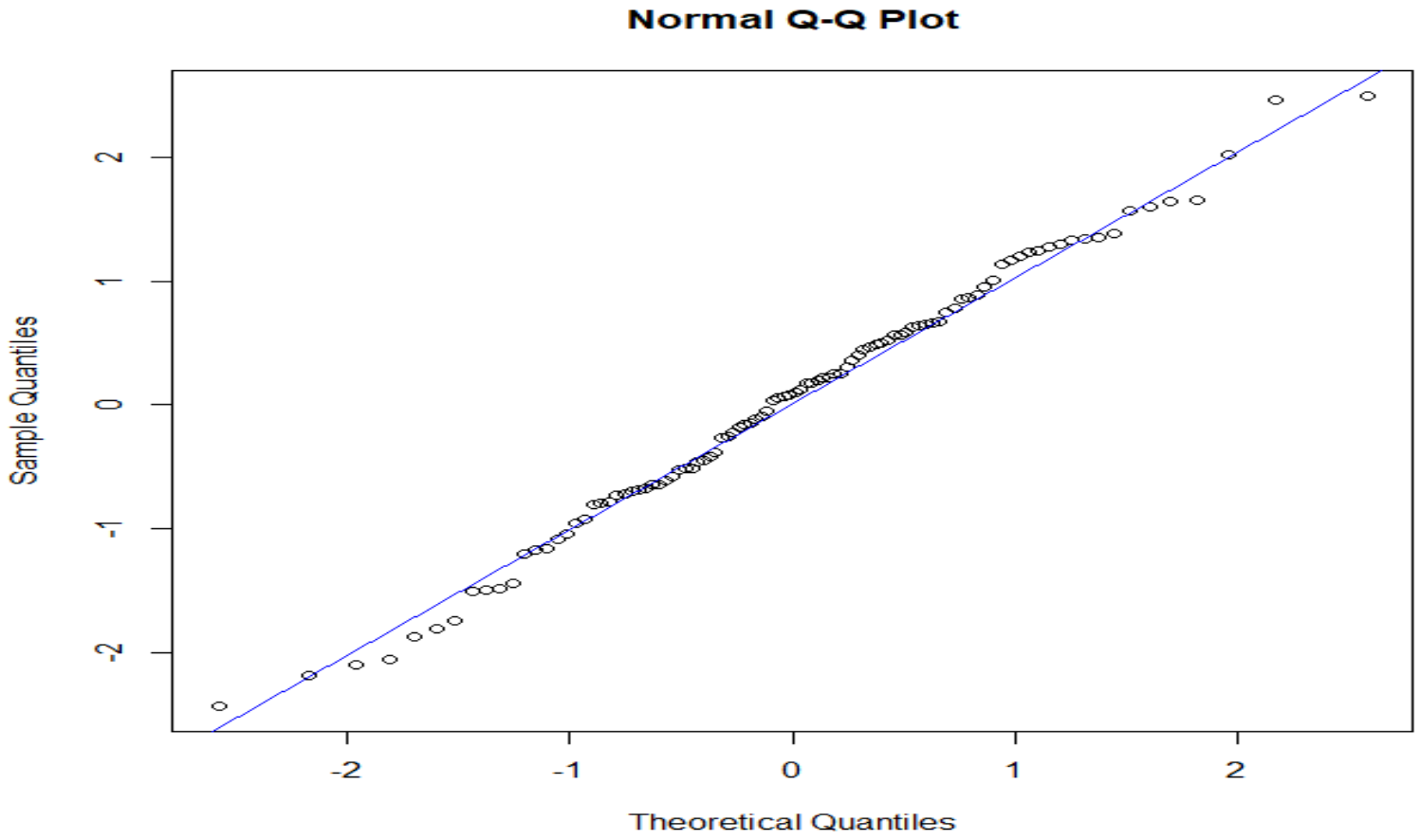
Quantile-Quantile Normal Plots

- qqnorm(x); qqline(x)
- qqplot(x,y) : visually check whether x and y are sampled from the same normal distribution

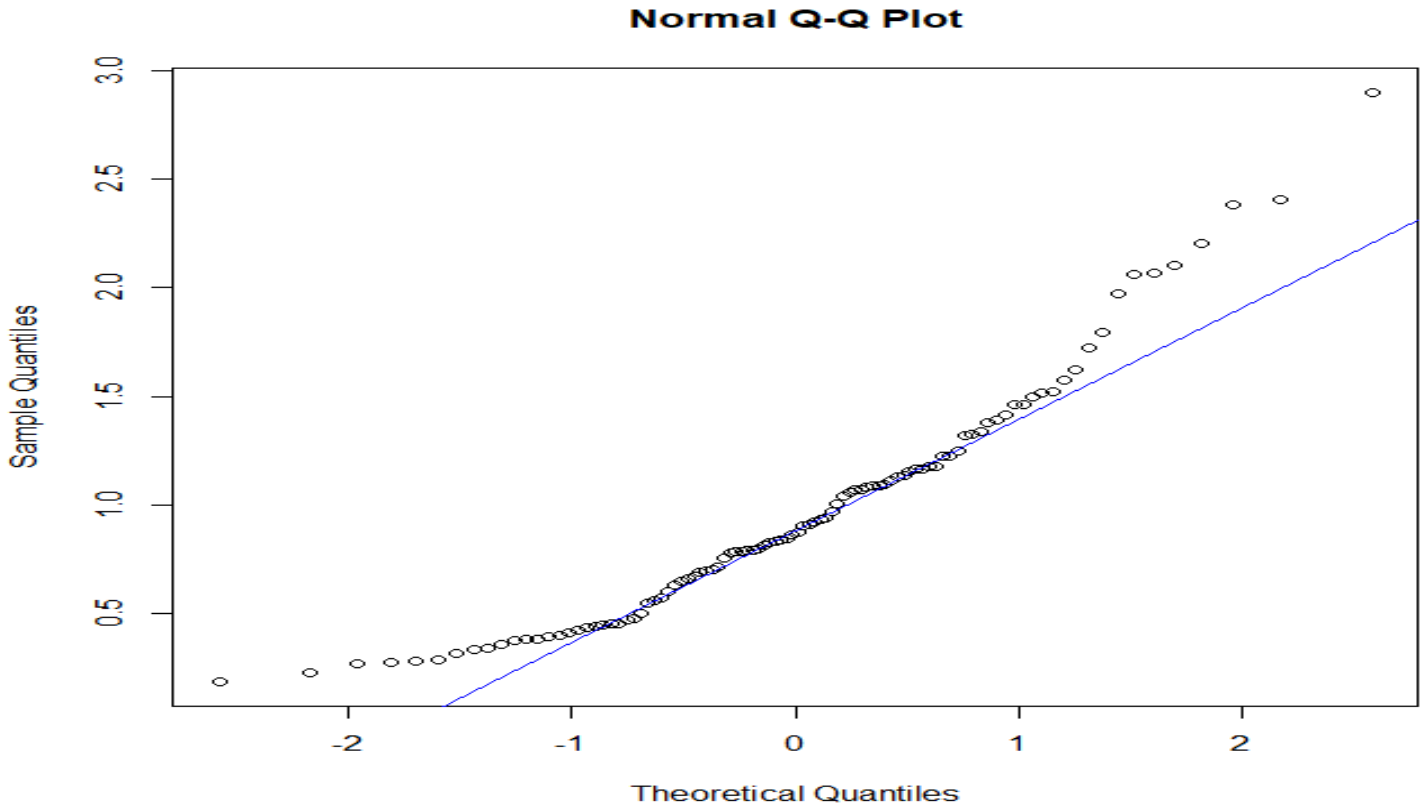
sim = rnorm(100) ; qqnorm(sim) ; qqline(sim, col="blue")



sim = rt(100, df=4) ; qqnorm(sim) ;
qqline(sim, col="blue")

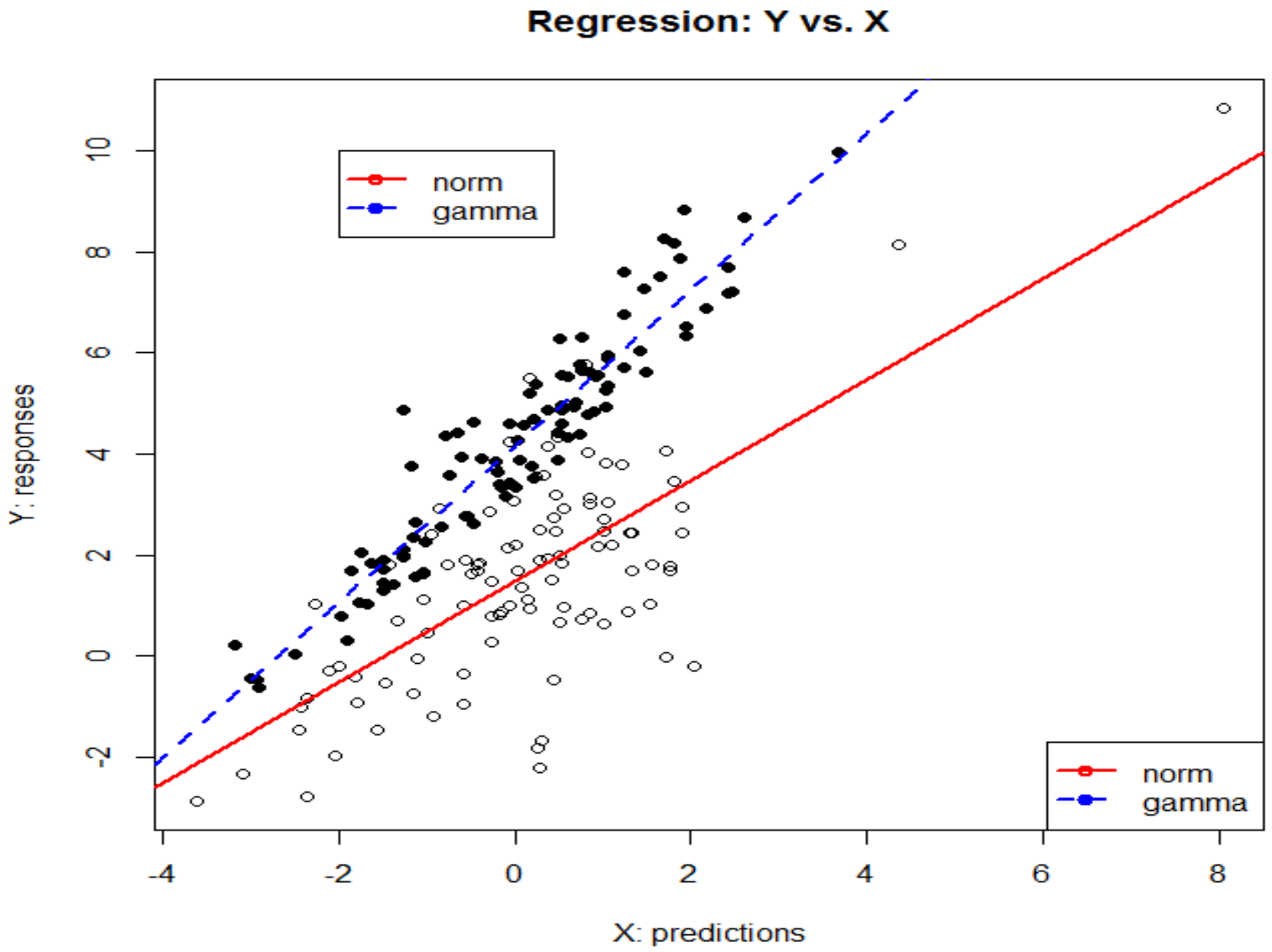


sim = rgamma(100, 3, 3) ; qqnorm(sim) ;
qqline(sim, col="blue")



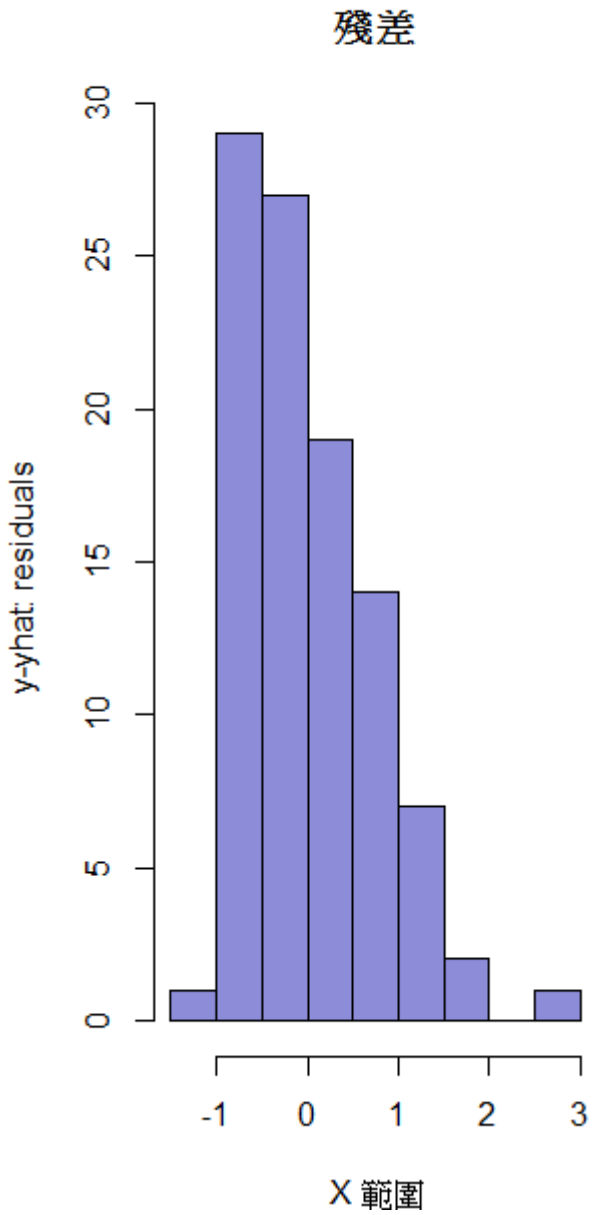
Other plot or diagram

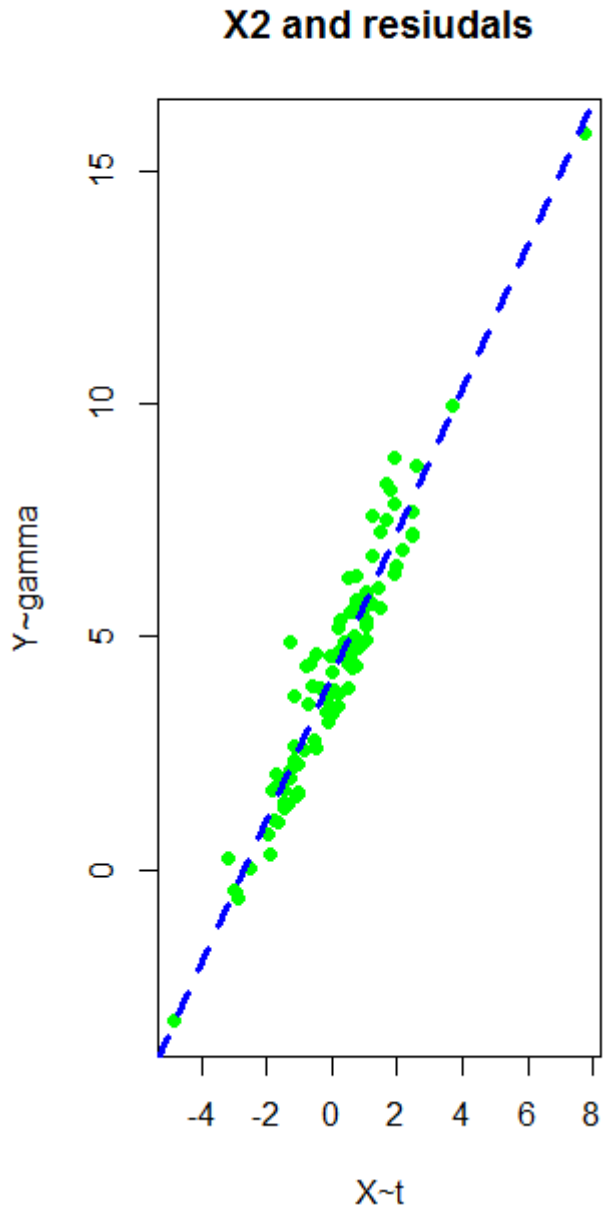
- plot(x, y, xlim=range(x), ylim=range(y), type="p", main, xlab, ylab, pch="1", col="blue")
- barplot(x, beside=FALSE, horiz=FALSE, legend=NULL)
- pie(x)
- …etc


Regression: Y vs. X
norm
gamma
norm
gamma
Y: responses
X: predictions
-4
-2
0
2
4
6
8
-2
0
2
4
6
8
10

```
set.seed(2018) ; n= 200
x = rt(n, df=4)     ; x1 = x[1:floor(n/2)] ; x2 = x[ (floor(n/2)+1) : n]
y1 = 2 + 1*x1 + rnorm( length(x1) , sd=2 )        ## 簡單迴歸，殘差為norm
y2 = 3 + 1.5*x2 + rgamma( length(x2) , 2, 2 )   ## 簡單迴歸，殘差為gamma

plot( x1, y1, type = “p”,main = “Regression: Y vs. X”,xlab = “X: predictions”,
ylab = “Y: responses”)  ## 建構主圖
points( x1, y1, pch = 1)  ; points( x2, y2, pch = 19)  ## 在主圖上添加點

abline(lm( y1 ~ x1 ), lty=1, lwd=2, col = “red” )   ##在主圖上畫簡單迴歸線
abline(lm( y2 ~ x2 ), lty=2, lwd=2, col = “blue”)  ##在主圖上畫簡單迴歸線
legend( -2, 10, c(“norm”,“gamma”), pch=c(1,19), lty=1:2, lwd=2, col=c(“red”,
“blue”))  ##在主圖某座標上添加敘述
legend( “bottomright”, c(“norm”,“gamma”), pch=c(1,19), lty=1:2, lwd=2,
col=c(“red”, “blue”)) ##在主圖某位置上添加敘述
```

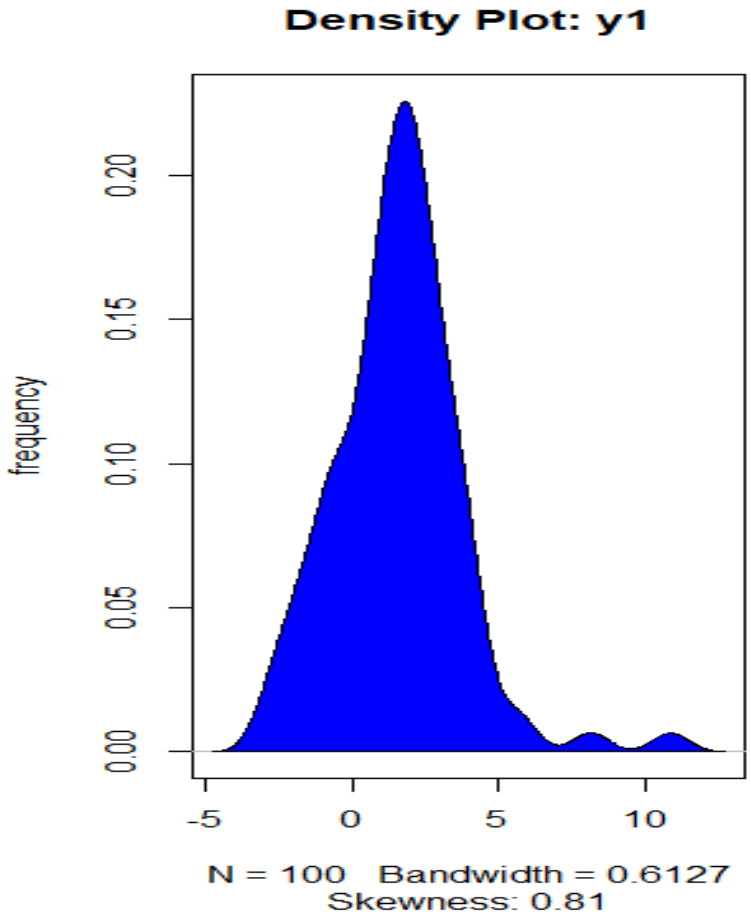
Regression: Histogram of residuals and Scatterplot

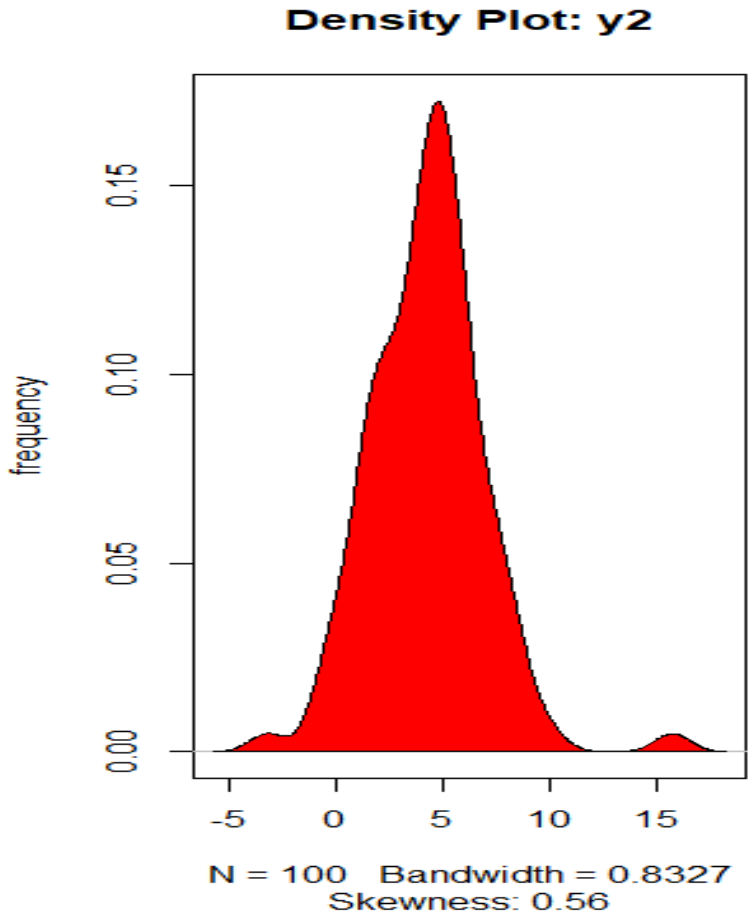




```
#Divide the screen in 1 line and 2 columns
par( mfrow=c(1,2), oma = c(0, 0, 2, 0) ) # 設定圖形的外邊界大小
# oma = c(bottom,left,top,right)
#設置圖形空白邊界行數，mar = c(bottom, left, top, right)
par(mar=c(4,4,4,2))

hist( lm( y2 ~ x2 )$residuals,  main="殘差" , col=rgb(0.1, 0.1, 0.7,0.5),  xlab="X
範圍" , ylab="y-yhat: residuals") # rgb( 紅，綠，藍，透明度 )

plot( x2 , y2 ,  main="X2 and resiudals" , pch=19 , cex=1 , col="green"  ,
xlab="X~t" , ylab="Y~gamma" )

abline(lm( y2 ~ x2 ), lty=2, lwd=3, col = "blue")

# add only ONE title
mtext(" Regression: Histogram of residuals and Scatterplot", outer = TRUE, cex =
1.5, font=4, col=rgb(0.7,0.1,0.1,0.7) )
```

Density plot





```
par(mfrow=c(1, 2))
plot( density(y1), main="Density Plot: y1",
ylab="frequency", sub=paste("Skewness:",
round(e1071::skewness(y1), 2)))
polygon( density(y1), col="blue")
# 畫y1的機率密度函數圖，曲線下面積令為藍色
plot( density(y2), main="Density Plot: y2",
ylab="frequency", sub=paste("Skewness:",
round(e1071::skewness(y2), 2)))
polygon( density(y2), col="red")
# 畫y2的機率密度函數圖，曲線下面積令為紅色
```

Linear regression with R

```
> summary(lm( y1 ~ x1 ))

Call:
lm(formula = y1 ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9657 -0.9176 -0.0049  0.8394  3.8439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4746     0.1520   9.699 5.45e-16 ***
x1            0.9973     0.1005   9.919  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.518 on 98 degrees of freedom
Multiple R-squared:  0.501,     Adjusted R-squared:  0.4959
F-statistic: 98.39 on 1 and 98 DF,  p-value: < 2.2e-16

> lm( y1 ~ x1 )

Call:
lm(formula = y1 ~ x1)

Coefficients:
(Intercept)           x1
     1.4746       0.9973

> ?lm
```

$\hat{\beta}_0, \hat{\beta}_1$

$y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

coefficients	a named vector of coefficients
residuals	the residuals, that is response minus fitted values.
fitted.values	the fitted mean values.
rank	the numeric rank of the fitted linear model.
weights	(only for weighted fits) the specified weights.
df.residual	the residual degrees of freedom.
call	the matched call.
terms	the terms object used.
contrasts	(only where relevant) the contrasts used.
xlevels	(only where relevant) a record of the levels of the factors used in fitting.
offset	the offset used (missing if none were used).
y	if requested, the response used.
x	if requested, the model matrix used.
model	if requested (the default), the model frame used.
na.action	(where relevant) information returned by model.frame on the special handling of NAs.

lm(y~x)$coefficients

Formula versus model

$\mathrm{lm}(y \sim x)$	$y \sim \beta_0 + \beta_1 x$
$\mathrm{lm}(y \sim -1 + x)$	$y \sim \beta_1 x$
$\mathrm{lm}(y \sim x + \mathrm{I}(x\text{^}2))$	$y \sim \beta_0 + \beta_1 x + \beta_2 x^2$
$\mathrm{lm}(y \sim x_1 + x_2)$	$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\mathrm{lm}(y \sim x_1 : x_2)$ $\qquad$ $y \sim \beta_0 + \beta_1 x_1 x_2$

$\mathrm{lm}(y \sim x_1 * x_2)$ $\qquad$ $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

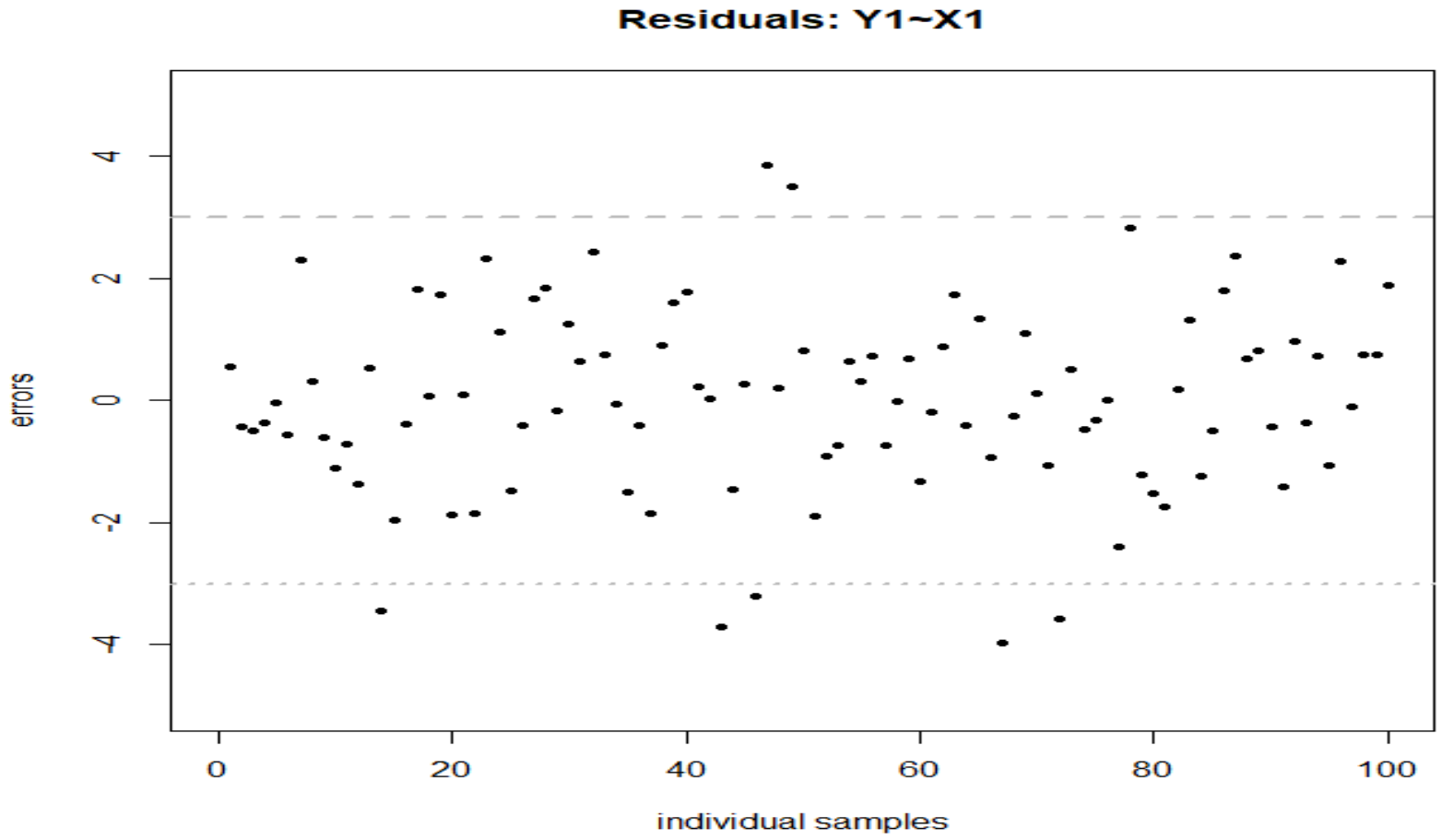
$\mathrm{lm}(y \sim (x_1 + x_2 + x_3) \wedge 2)$

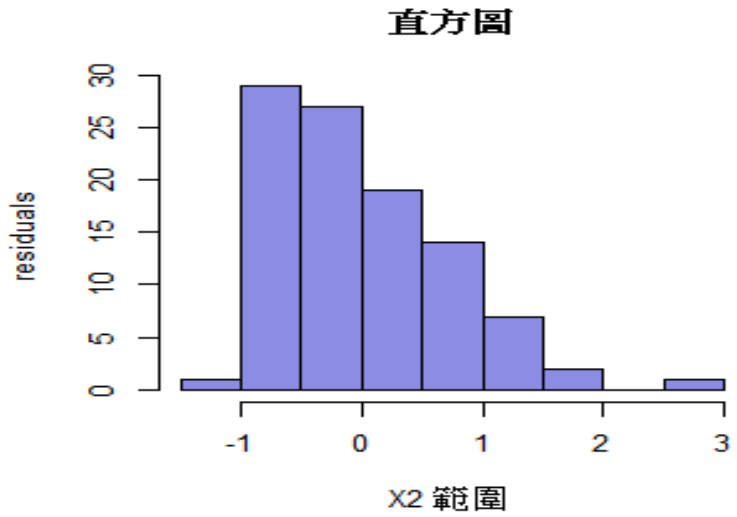
$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$

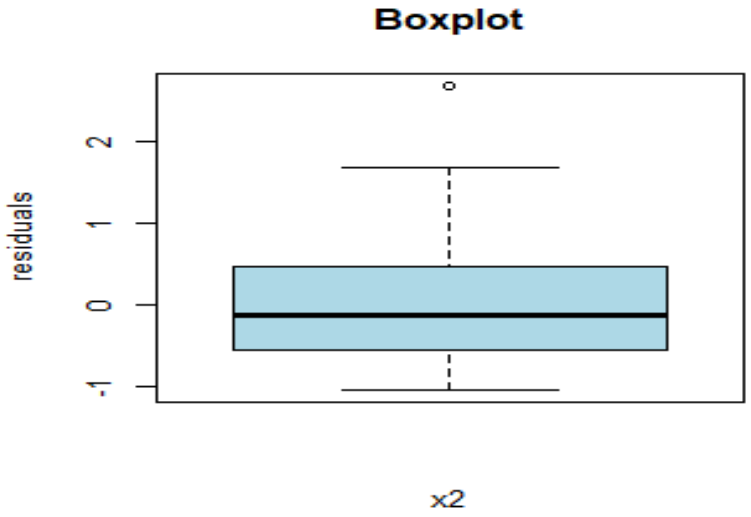
ANOVA

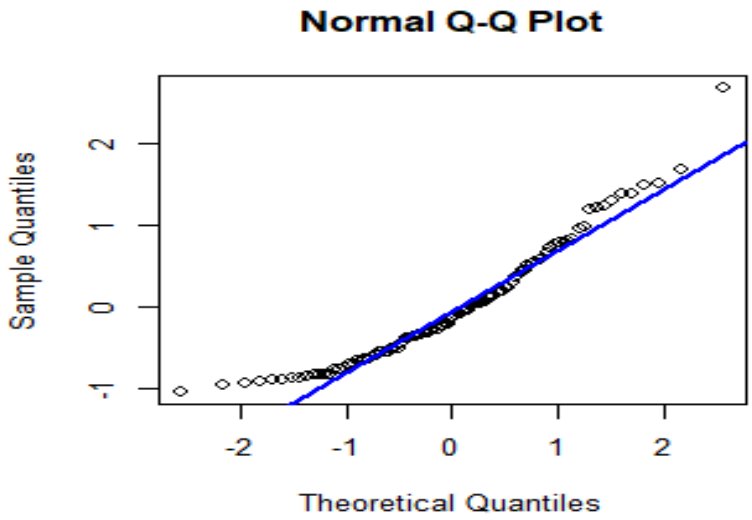
```
> anova(lm( y2 ~ x2 ))
Analysis of Variance Table

Response: y2
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 640.86  640.86  1198.5 < 2.2e-16 ***
Residuals 98  52.40    0.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

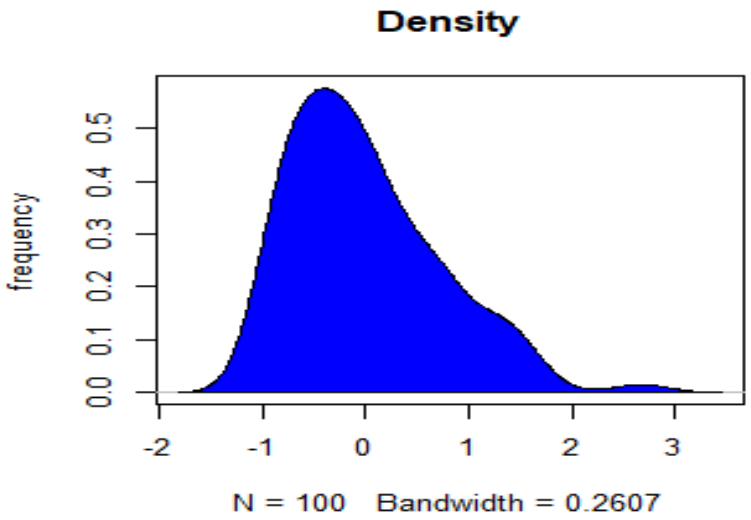
Residuals Plot




直方圖
residuals
0
5
10
15
20
25
30
-1
0
1
2
3
X2 範圍

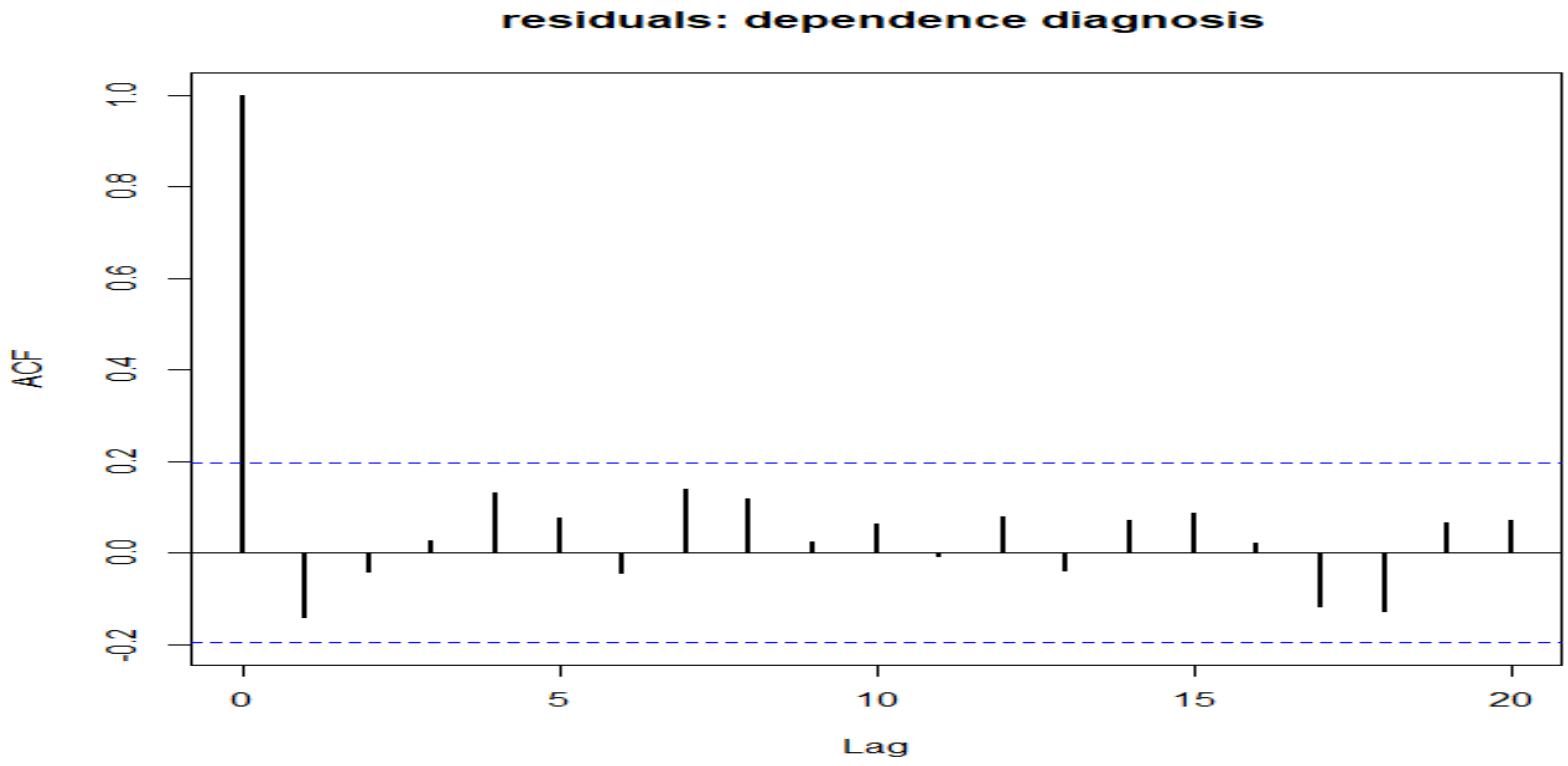

Boxplot
residuals
-1
0
1
2
x2


Normal Q-Q Plot
Sample Quantiles
-1
0
1
2
-2
-1
0
1
2
Theoretical Quantiles


Density
frequency
0.0
0.1
0.2
0.3
0.4
0.5
-2
-1
0
1
2
3
N = 100 Bandwidth = 0.2607

```
par(mfrow=c(2, 2)) ; error2 = lm( y2 ~
x2 )$residuals
hist( error2,  main="直方圖" , col=rgb(0.1, 0.1,
0.8,0.5),  xlab="X2 範圍" , ylab="residuals") #
rgb( 紅，綠，藍，透明度 )
boxplot( error2, xlab = "x2", ylab = "residuals", main
= "Boxplot", varwidth = TRUE, col = "lightblue")
qqnorm(error2) ; qqline(error2, col="blue", lwd = 2)
plot( density(error2), main="Density",
ylab="frequency" )
polygon( density(error2), col="blue")
```

AutoCorrelation Function (ACF)



acf(error2, main=" residuals: dependence diagnosis" , lwd = 2)