Lab Session-2: Normal Distribution (Uni & Multivariate)

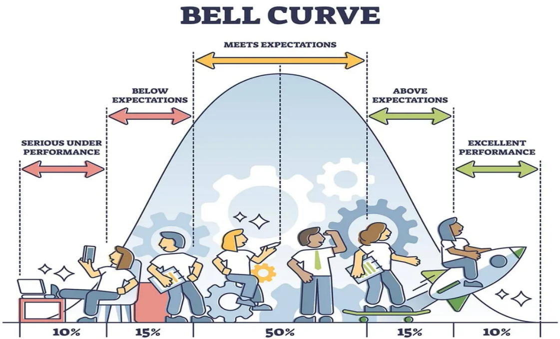
MATH350 – Statistical Inference

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty, Ph.D. from ISI Kolkata.
Assistant Professor in Statistics at Sorbonne University.
tanujit.chakraborty@sorbonne.ae
Course Webpage: https://www.ctanujit.org/SI.html
Course for BSc Mathematics and Data Science Students.

Normality & Beyond Normality

Normality is a paved road. It is easy to walk but no flowers grow on it. — Vincent Van Gogh.



By Dr. Saul McLeod (2019)

Few Famous Quotations

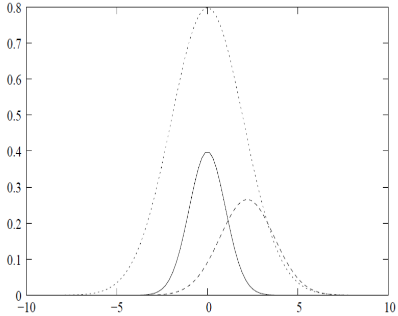
Normality is a myth; there never was, and there never will be a normal distribution — Roy C. Geary (1947; Biometrika, vol. 34, 248).

Everybody believes in the exponential law of errors (the normal distribution), the experimenters, because they think it can be proved by mathematicians; and the mathematicians, because they believe that it has been established by observations — E.T. Whittaker and G. Robinson (1967).

... the statisticians knows ... that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false he can often derive results which match to a useful approximation, those found in real world — George W. Box (1976, Journal of American Statistical Association, vol. 71, 791-799).

A random variable X is said to be normally distributed with mean μ and variance σ^2, if the probability density function of X is the following (for $-\infty < \mu < \infty$ and $\sigma > 0$)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ -\infty < x < \infty$$



Probability Density Function of Normals

Galton Board

- Sir Francis Galton, Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.

- It specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.

Picture of Galton Board

Gambling Question: A 17th century gambler, the Chevalier de Mere, asked Pascal for an explanation of his unexpected losses in gambling.

The famous correspondence between Pascal and Fermat was instigated in 1654, and they were mainly interested to calculate the following binomial sum:

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^{n-k}$$

The problem was not difficult when n is small.

Within few years the following problem arises in a sociological study, where the following computation was necessary:
$n = 11,429, i = 5745, j = 6128$

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^{n-k}$$

Original Problem: The problem is to test the hypothesis that male and female births are equally likely against the actual birth in London over 82 years from 1629 - 1710. It is observed that the relative number of male births varies from a low of $7765/15,448 = 0.5027$ in 1703 to a high of $4748/8855 = 0.5362$ in 1661. Given that 11,429 is the average number of births in London over 82 years, and 5745 and 6128 are two limits.

Using the following recurrence relation

$$\binom{n}{x+1} = \binom{n}{x}\binom{n-x}{x+1}$$

and some involved rational approximation it has been obtained

$$P(5747 \leq X \leq 6128 \mid p = 1/2) = \sum_{i=5745}^{6128} \binom{11,429}{i} \left(\frac{1}{2}\right)^{i}$$
$$\approx 0.292$$

Solution

Using the following recurrence relation

$$\binom{n}{x+1} = \binom{n}{x}\binom{n-x}{x+1}$$

and some involved rational approximation it has been obtained

$$P(5747 \leq X \leq 6128 \mid p = 1/2) = \sum_{i=5745}^{6128} \binom{11,429}{i} \left(\frac{1}{2}\right)^i$$
$$\approx 0.292$$

De Moivre began the search for this approximation in 1721 , and in 1733 it has been proved that

$$\binom{n}{\frac{n}{2}+x}\left(\frac{1}{2}\right)^n \approx \frac{2}{\sqrt{2\pi n}} e^{-2x^2/n}$$

and

$$\sum_{|x-n/2|\leq a} \binom{n}{x}\left(\frac{1}{2}\right)^n \approx \frac{4}{\sqrt{2\pi}} \int_0^{a/\sqrt{n}} e^{-2y^2} dy.$$

Normal Approximation

Eventually using the second approximation one gets

$$\sum_{k=i}^{j} \binom{n}{k} p^k (1-p)^k \approx \Phi\left(\frac{j-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{i-np}{\sqrt{np(1-p)}}\right)$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

which is the cumulative distribution function (CDF) of the standard normal distribution.

Error Modeling

Gauss (1809) made the following assumptions and deduce the normal distribution as an error distribution:

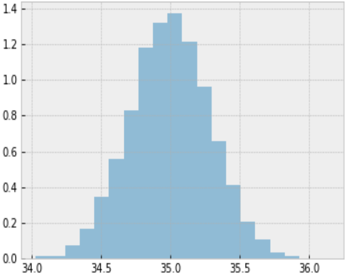
1. Small errors are more likely than large errors.
2. For any real numbers ϵ, the likelihood of errors of magnitudes ϵ and $-\epsilon$ are equal.
3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

To read more about the evolution of normal distribution: Saul Stahl (2006), "The evolution of normal distribution", Mathematics Magazine, vol. 79, no. 2, 96 - 113.

Lindeberg-Levy CLT:

Suppose $\{X_1, X_2, \cdots\}$ is a sequence of independent identically distributed random variables with mean μ and variance $\sigma^2 < \infty$, then as $n \to \infty$

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \to N(0,1)$$



CLT in Practice

Some Drawbacks

What will happen if the data indicate that the parent distribution

1. is not symmetric?
2. is heavy tail?
3. is not unimodal?

What will happen if error distribution is not normal during regression modeling?

Possible Remedy

In Distribution Theory:

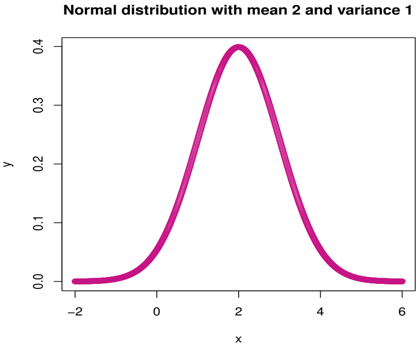
1. Skew Normal Distribution (A Azzalini, Scandinavian Journal of Statistics 1985)
2. Power Normal Distribution (RD Gupta, Test 2008)
3. Geometric Skew-Normal Distribution (D Kundu, Sankhya 2014), etc.

In Regression Theory:

1. Box-Cox Transformation (Box, Cox, JRSS Series-B 1964)
2. Generalized linear model (Nelder, Wedderburn, JRSS Series-A 1972)
3. Semiparametric and Nonparametric Approaches (see ESLR/ISLR Book), etc.

Univariate normal with mean 2 and variance 1

x <- seq(-2, 6, by = .01)
y <- dnorm(x, mean = 2, sd = 1)
plot(x,y,col = "mediumvioletred")



Functions for statistical distributions in R

Normal	**Multivariate Normal**	**t**	**Multivariate t**
rnorm	rmvnorm	rt	rmvt
dnorm	dmvnorm	dt	dmvt
pnorm	pmvnorm	pt	pmvt
qnorm	qmvnorm	qt	qmvt

Functions for statistical distributions in R

Normal	Multivariate Normal	t	Multivariate t
rnorm	rmvnorm	rt	rmvt
dnorm	dmvnorm	dt	dmvt
pnorm	pmvnorm	pt	pmvt
qnorm	qmvnorm	qt	qmvt

The first letter denotes

- **r** for “simulation"
- **d** for “density"
- **p** for “probability"
- **q** for “quantile"

Followed by the distribution name

- **norm**
- **mvnorm**
- **t**

The rmvnorm function

install.package (“mvtnorm”)
library (mvtnorm)
rmvnorm (n, mean , sigma)

Parameters need to be specified:

- **n** the number of samples
- **mean** the mean of the distribution
- **sigma** the variance-covariance matrix

Generate 1000 samples from a 3 dimensional normal with

$$\mu = \begin{pmatrix} 1 \\ 2 \\ -5 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

```
mu1 <- c (1, 2, -5)
sigma1 <- matrix ( c ( 1,1,0,1,2,0,0,0,5 ), 3,3 )
set.seed (34)
sim_mv = rmvnorm (n = 1000, mean = mu1, sigma = sigma1)
library ("corrplot")
corrplot (cor (sim_mv), method = "ellipse")
```

Plot of generated samples

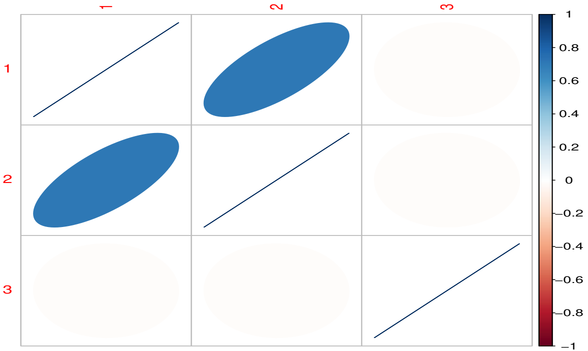


Figure: Correlation plot of the generated sample

Density using dmvnorm

install.package ("mvtnorm")
library (mvtnorm)
dmvnorm (x, mean , sigma)

Parameters need to be specified:

- **x** can be a vector or matrix
- **mean** the mean of the distribution
- **sigma** the variance-covariance matrix

Density using dmvnorm

Compute the density at $(0, 0)$ from normal distribution with mean and variance-covariance matrix as

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

```
mu1 <- c (1, 2)
sigma1 <- matrix ( c ( 1, .5, .5, 2 ) , 2 )
dmvnorm ( x = c ( 0, 0 ), mean = mu1, sigma = sigma1)
```

Output: *0.03836759*

Density at multiple points using dmvnorm

Compute the density at $x = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$ from normal distribution with mean and variance-covariance matrix as

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

Density at multiple points using dmvnorm

Compute the density at $x = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$ from normal distribution with mean and variance-covariance matrix as

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

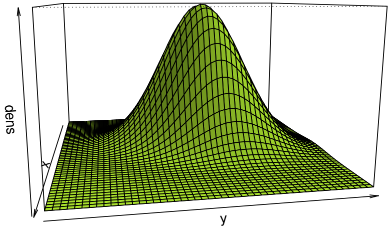
```
x <- rbind ( c ( 0, 0 ), c ( 1, 1 ), c ( 0, 1 ) )
mu1 <- c (1, 2)
sigma1 <- matrix ( c ( 1, .5, .5, 2 ) , 2 )
dmvnorm ( x = c ( 0, 0 )x, mean = mu1, sigma = sigma1)
```

Output: 0.03836759 0.09041010 0.06794114

Steps:

- Create grid of **x** and **y** coordinates
- Calculate density on grid
- Convert densities into a matrix
- Create perspective plot using

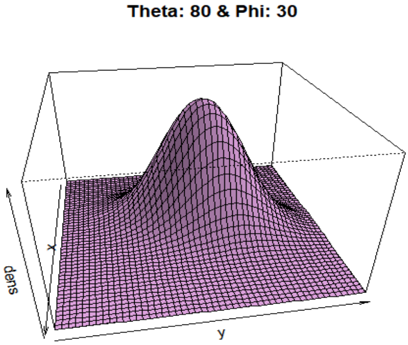
persp() function



Plotting bivariate densities

```
d <- expand.grid ( seq ( -3, 6, length.out = 50 ), seq( -3, 6, length.out =
50 ) )
dens1 <- dmvnorm ( as.matrix ( d ), mean = c ( 1, 2 ), sigma = matrix ( c
( 1, .5, .5, 2 ), 2 ) )
dens1 <- matrix(dens1, nrow = 50 )
persp(dens1, theta = 80, phi = 30, expand = 0.6, shade = 0.2,
col = "plum1" , xlab = "x" , ylab = "y" , zlab = "dens")
```

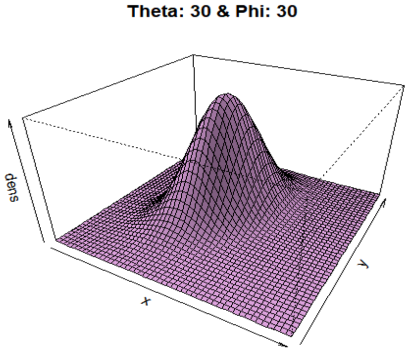
Code for plotting bivariate densities



persp() with theta = 30, phi = 30

```
d <- expand.grid ( seq ( -3, 6, length.out = 50 ), seq( -3, 6, length.out =
50 ) )
dens1 <- dmvnorm ( as.matrix ( d ), mean = c ( 1, 2 ), sigma = matrix ( c
( 1, .5, .5, 2 ), 2 ) )
dens1 <- matrix(dens1, nrow = 50 )
persp(dens1, theta = 30, phi = 30, expand = 0.6, shade = 0.2,
col = "plum1" , xlab = "x" , ylab = "y" , zlab = "dens", main = "Theta:
30  Phi: 30")
```

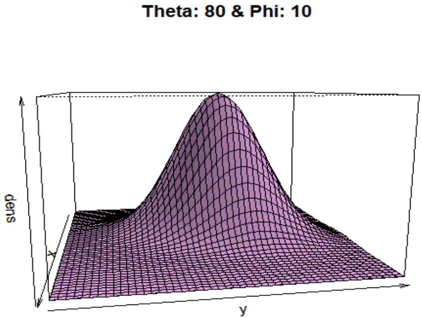
Plotting bivariate density



persp() with theta = 80, phi = 10

```
d <- expand.grid ( seq ( -3, 6, length.out = 50 ), seq( -3, 6, length.out =
50 ) )
dens1 <- dmvnorm ( as.matrix ( d ), mean = c ( 1, 2 ), sigma = matrix ( c
( 1, .5, .5, 2 ), 2 ) )
dens1 <- matrix(dens1, nrow = 50 )
persp(dens1, theta = 80, phi = 10, expand = 0.6, shade = 0.2,
col = "plum1" , xlab = "x" , ylab = "y" , zlab = "dens", main = "Theta:
80  Phi: 10")
```

Plotting bivariate density



Calculating CDF and inverse CDF

Compute the probability at $x \leq 200$ where x is distributed as a normal distribution with mean 210 and variance 100.

pnorm (200, mean = 210, sd = 10)

Output: *0.1586553*

What is the x_0 such that the cumulative probability at x_0 is 0.95?

qnorm (p = 0.95, mean = 210, sd = 10)

Output: *226.4485*

Cumulative distribution using pmvnorm

Bivariate CDF at x = 2 and y = 4 for a normal with

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

```
mu1 <- c ( 1, 2 )
sigma1 <- matrix ( c ( 1, 0.5, 0.5, 2 ) , 2 )
pmvnorm ( upper = c ( 2, 4 ) , mean = mu1, sigma = sigma1)
```

Output:

```
0.79
attr(,"error")
1e-15
attr ( "msg" )
"Normal Completion"
```

Probability of $1 < x < 2$ and $2 < y < 4$ for a normal with

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

mu1 <- c (1, 2)
sigma1 <- matrix (c (1, 0.5, 0.5, 2) , 2)
pmvnorm (lower = c(1, 2), upper = c(2, 4), mean = mu1, sigma = sigma1)
Output: *[1] 0.163*

```
sigma1 <- diag ( 2 )
sigma1
```

Output: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

```
qmvnorm ( p = 0.95, sigma = sigma1, tail = "both")
```

Output:

```
$quantile
2.24
$f.quantile
-1.31e-06
attr(, "message")
"Normal Completion"
```

Checking normality of multivariate data

Why check normality?
Classical statistical techniques that assume univariate or multivariate normality:

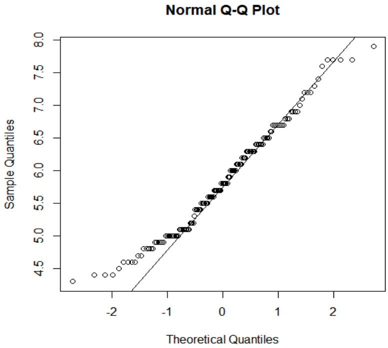
- Multivariate regression
- Discriminant analysis
- Model-based clustering
- Principal component analysis (PCA)
- Multivariate analysis of variance (MANOVA)

Univariate normality tests

Check whether “Sepal.Length" attribute of iris dataset in R follows a normal distribution.

```
qqnorm ( iris [, 1] )
qqline ( iris [, 1] )
```

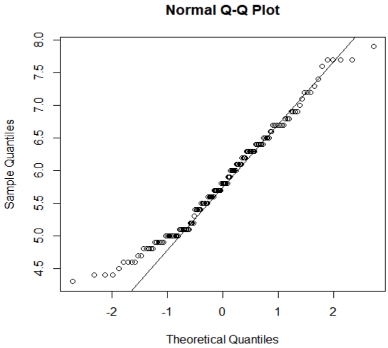
- If the values lie along the reference line the distribution is close to normal.



Univariate normality tests

Check whether “Sepal.Length" attribute of iris dataset in R follows a normal distribution.

- If the values lie along the reference line the distribution is close to normal.
- Deviation from the line might indicate the following :
  - heavier tails
  - skewness
  - outliers
  - clustered data



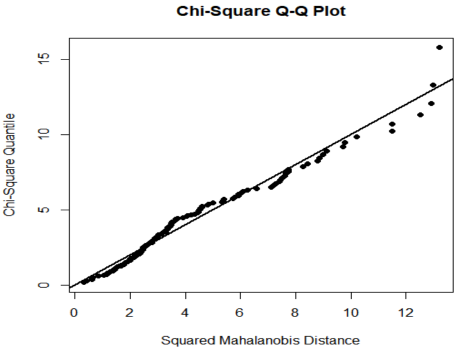
- Multivariate normality tests by
  - Mardia
  - Henze-Zirkler
  - Royston
- Graphical appoaches
  - chi-square Q-Q
  - perspective
  - contour plots

install.packages ("MVN")
library (MVN)
mvn (iris [, 1:4] , subset = NULL, mvnTest = "mardia")

```
$multivariateNormality
             Test          Statistic              p value Result
1 Mardia Skewness    67.430508778062 4.75799820400869e-07     NO
2 Mardia Kurtosis -0.230112114481001    0.818004651478012    YES
3             MVN               <NA>                 <NA>     NO

$univariateNormality
              Test     Variable Statistic   p value Normality
1 Anderson-Darling Sepal.Length    0.8892  0.0225      NO
2 Anderson-Darling Sepal.Width     0.9080  0.0202      NO
3 Anderson-Darling Petal.Length    7.6785  <0.001      NO
4 Anderson-Darling Petal.Width     5.1057  <0.001      NO

$Descriptives
               n     Mean   Std.Dev Median Min Max 25th 75th       Skew   Kurtosis
Sepal.Length 150 5.843333 0.8280661   5.80 4.3 7.9  5.1  6.4  0.3086407 -0.6058125
Sepal.Width  150 3.057333 0.4358663   3.00 2.0 4.4  2.8  3.3  0.3126147  0.1387047
Petal.Length 150 3.758000 1.7652982   4.35 1.0 6.9  1.6  5.1 -0.2694109 -1.4168574
Petal.Width  150 1.199333 0.7622377   1.30 0.1 2.5  0.3  1.8 -0.1009166 -1.3581792
```

Iris data is not multivariate normal

mvn (iris [, 1:4], subset = NULL, mvnTest = "mardia", multivariatePlot = "qq")



install.packages ("MVN")
library (MVN)
mvn (iris [, 1:4] , subset = NULL, mvnTest = "hz")

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 2.336394       0  NO

$univariateNormality
              Test     Variable Statistic   p value Normality
1 Anderson-Darling Sepal.Length    0.8892  0.0225      NO
2 Anderson-Darling Sepal.Width     0.9080  0.0202      NO
3 Anderson-Darling Petal.Length    7.6785  <0.001      NO
4 Anderson-Darling Petal.Width     5.1057  <0.001      NO

$Descriptives
               n     Mean   Std.Dev Median Min Max 25th 75th       Skew   Kurtosis
Sepal.Length 150 5.843333 0.8280661   5.80 4.3 7.9  5.1  6.4  0.3086407 -0.6058125
Sepal.Width  150 3.057333 0.4358663   3.00 2.0 4.4  2.8  3.3  0.3126147  0.1387047
Petal.Length 150 3.758000 1.7652982   4.35 1.0 6.9  1.6  5.1 -0.2694109 -1.4168574
Petal.Width  150 1.199333 0.7622377   1.30 0.1 2.5  0.3  1.8 -0.1009166 -1.3581792
```

Iris data is not multivariate normal

Testing multivariate normality by species

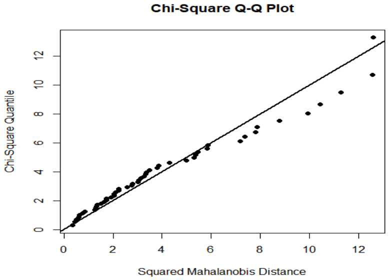
install.packages ("MVN")
library (MVN)
mvn (iris [iris $ Species == "setosa", 1:4], subset = NULL, mvnTest = "mardia")

```
$multivariateNormality
             Test        Statistic           p value Result
1 Mardia Skewness 25.6643445196298 0.177185884467652    YES
2 Mardia Kurtosis 1.29499223711605 0.195322907441935    YES
3             MVN             <NA>              <NA>    YES

$univariateNormality
              Test     Variable Statistic   p value Normality
1 Anderson-Darling Sepal.Length    0.4080  0.3352      YES
2 Anderson-Darling Sepal.Width     0.4910  0.2102      YES
3 Anderson-Darling Petal.Length    1.0073  0.0108      NO
4 Anderson-Darling Petal.Width     4.7148  <0.001      NO

$Descriptives
              n  Mean   Std.Dev Median Min Max 25th  75th       Skew   Kurtosis
Sepal.Length 50 5.006 0.3524897    5.0 4.3 5.8  4.8 5.200 0.11297784 -0.4508724
Sepal.Width  50 3.428 0.3790644    3.4 2.3 4.4  3.2 3.675 0.03872946  0.5959507
Petal.Length 50 1.462 0.1736640    1.5 1.0 1.9  1.4 1.575 0.10009538  0.6539303
Petal.Width  50 0.246 0.1053856    0.2 0.1 0.6  0.2 0.300 1.17963278  1.2587179
```

Data is multivariate normal

Checking QQ plot by species

install.packages ("MVN")
library (MVN)
mvn (iris [iris $ Species == "setosa", 1:4], subset = NULL,
mvnTest = "mardia", multivariatePlot = "qq")



Data is multivariate normal