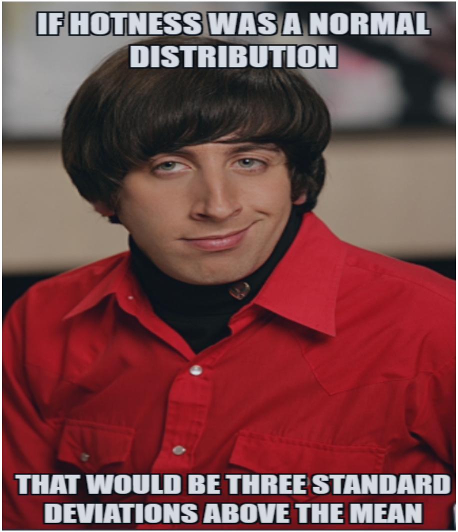
Lab Session-2: Normality and Beyond

MATH350 – Statistical Inference

STATISTICS + MACHINE LEARNING + DATA SCIENCE

Dr. Tanujit Chakraborty
Assistant Professor in Statistics at Sorbonne University
tanujit.chakraborty@sorbonne.ae
Course Webpage: https://www.ctanujit.org/SI.html
R Code: https://github.com/tanujit123/MATH350

Let's Start the Journey to Normality

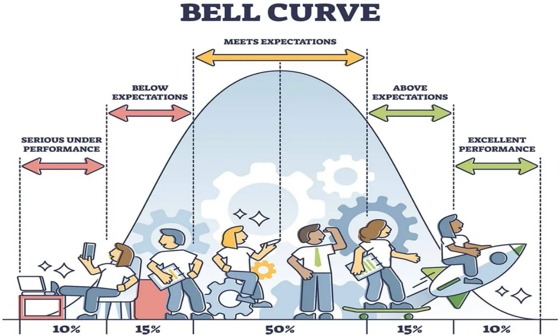


Topics for Today's Session

- Normality: A Brief History
- Univariate Normal Distribution
- Drawbacks and Skew Normal Distribution
- Multivariate data: Iris Data
- Data Visualisation

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).

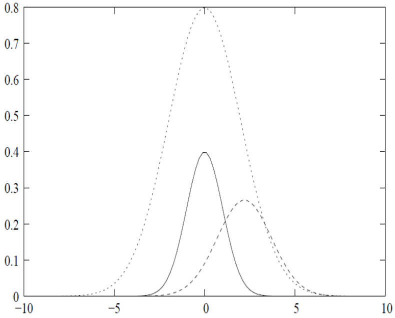
Everybody believes in the exponential law of errors (the normal distribution), the experimenters, because they think it can be proved by mathematicians; and the mathematicians, because they believe that it has been established by observations — E.T. Whittaker and G. Robinson (1967).

... 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).

Normal Distribution

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:
$n = 11,429, i = 5745, j = 6128$

$$\sum_{k=i}^{j} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

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.*

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

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

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

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

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

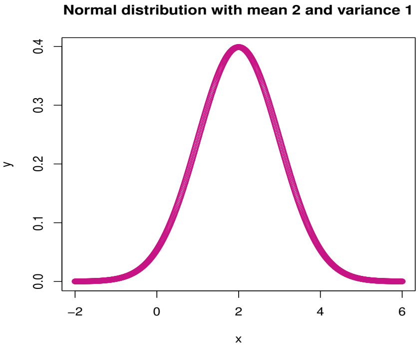
Gauss (1809) made the following assumptions and deduce the normal distribution as an error distribution:

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.

Univariate normal distribution

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")



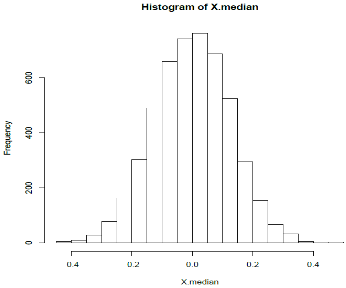
Simulating a sample median

Simulating a sample median. Let $X_1, \ldots, X_{99} \overset{IID}{\sim} \mathcal{N}(0,1)$. *The sample median is the 50th largest value among* $X_1, \ldots, X_{99}$. *Compute the sample medians from 5000 simulations of* $X_1, \ldots, X_{99}$. *What is the mean of these 5000 sample medians? What is their standard deviation? Plot a histogram of the 5000 values - what does the sampling distribution of the sample median look like?*

Solution

R Code:

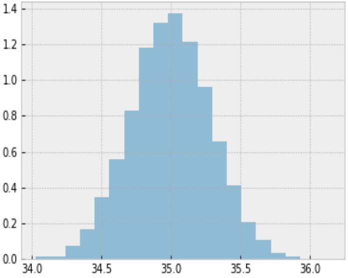
```
X.median = numeric(5000)
for(i in 1:5000) {
X = rnorm(99, mean = 0, sd = 1)
X.median[i] = median(X)
}
print(mean(X.median))
print(sd(X.median))
hist(X.median)
```



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

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.

Skew Normal Distribution

Goal:

1. Generate a non-symmetric class of distributions which have support on the whole real line.
2. Normal distribution is a special member.
3. It should not have too many parameters.

Construction:

- Suppose X and Y are two independent standard normal random variables, and λ is any real number. Therefore

$$P(X < \lambda Y) = P(X - \lambda Y < 0) = \frac{1}{2}$$

as $X - \lambda Y$ is a normal random variable with mean 0 , and variance $1 + \lambda^2$.

- On the other hand

$$P(X < \lambda Y) = \int_{-\infty}^{\infty} \Phi(\lambda y)\phi(y)dy$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{and} \quad \Phi(x) = \int_{-\infty}^{x} \phi(u)du.$$

- Therefore,

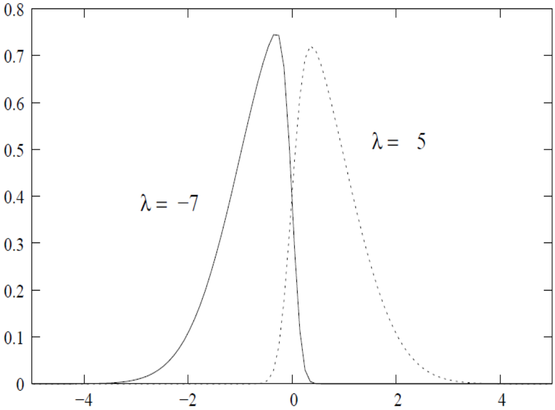
$$\frac{1}{2} = \int_{-\infty}^{\infty} \Phi(\lambda y)\phi(y)dy.$$

Since $\Phi(\lambda y)\phi(y) \geq 0$, the function

$$f(x;\lambda) = 2\phi(x)\Phi(\lambda x)$$

is a proper probability density function, and it is called skew-normal probability density function with parameter λ and we will denote it by $\mathrm{SN}(\lambda)$.

Shapes of the PDF SN



(1) The $\mathrm{SN}(0)$ density is the $\mathrm{N}(0,1)$ density.
(2) As $\lambda \to \infty$,

$$f(x;\lambda) \to \sqrt{\frac{2}{\pi}} e^{-x^2/2}; \quad x > 0$$

(3) If Z is a $\mathrm{SN}(\lambda)$ random variable, then $-Z$ is a $\mathrm{SN}(-\lambda)$ random variable.
(4) The PDF of a SN (λ) random variable is unimodal.
(5) If Z is $SN(\lambda)$ then Z^2 is χ_1^2.

For data analysis purposes three-parameter skew normal distribution can be easily defined with the probability density function as follows:

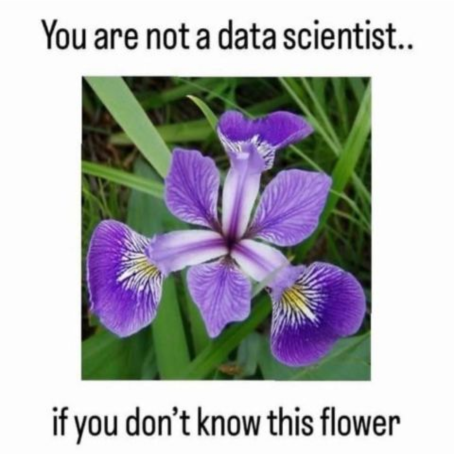
$$f(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{\lambda(x-\mu)}{\sigma}\right)$$

Structure of Multivariate Data

- Rectangular in shape - organized by rows and columns
  - Rows represent observations
  - Columns represent variables
- May or may not include:
  - Row names or numbers
  - Column headers
- Possible missing data

Example : Iris Data

This is perhaps the best known database to be found in the ML literature created by R.A. Fisher. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.



Multivariate data examples

Iris data from *'datasets'* package in R.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

The Iris dataset comprises 150 observations (rows) on the following 5 variables (columns)

- *Sepal.Length* - length (in cm) of the flower's sepal.
- *Sepal.Width* - width (in cm) of the flower's sepal.
- *Petal.Length* - length (in cm) of the flower's petal.
- *Petal.Width* - width (in cm) of the flower's petal.
- *Species* - categorical variable represents the category of the flower.

Exploring Iris data

Reading data

```
data (iris)
head (iris, n= 4)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa

Check the dataset dimension

```
dim (iris)
```

150 5

Extract the column names

names (iris)

"Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

Access Sepal length and Sepal width columns of observations 8 to 10

iris [8:10, 1:2]

Sepal.Length	Sepal.Width
5.0	3.4
4.4	2.9
4.9	3.1

Exploring Iris data

Check the data types

str (iris)

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2
 $ Species     : Factor w/ 3 levels "setosa","versicolor"
```

Reassign factor labels

Re-code the factors “setosa" , “versicolor", and “virginica" of *Species* variable to "1", "2", and "3"

library (car)
iris$Species <- recode (iris$Species, "'setosa' = 1; 'versicolor' = 2; 'virginica' = 3")

Calculate mean

```
colMeans (iris [, 1:4])
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.843333	3.057333	3.758000	1.199333

Functions that calculate means by subgroups

- ✓ *by* ()
- ✓ *aggregate* ()

Calculating the group mean

by (data = iris [,1:4], INDICES = iris$Species, FUN = colMeans)

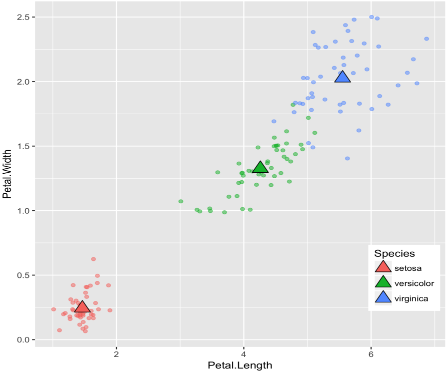
```
iris$Species: 1
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.006        3.428        1.462        0.246
------------------------------------------------------------
iris$Species: 2
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.936        2.770        4.260        1.326
------------------------------------------------------------
iris$Species: 3
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       6.588        2.974        5.552        2.026
```

aggregate (. ∼ Species, iris, mean)

```
  Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1       1        5.006       3.428        1.462       0.246
2       2        5.936       2.770        4.260       1.326
3       3        6.588       2.974        5.552       2.026
```

Interpretation of Means

Species	Petal.Length	Petal.Width
setosa	1.46	0.244
versicolor	4.26	1.326
virginica	5.55	2.026



var (iris[, 1:4])

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

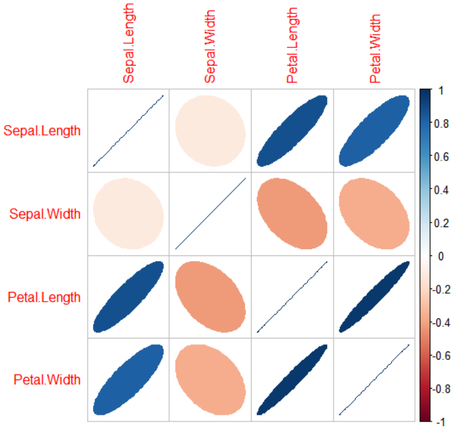
cor (iris[, 1:4])

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Visualization of correlation matrix

corrplot function to visualize correlation plot

```
library (corrplot)
corrplot (cor (iris [ , 1:4]), method = "ellipse")
```

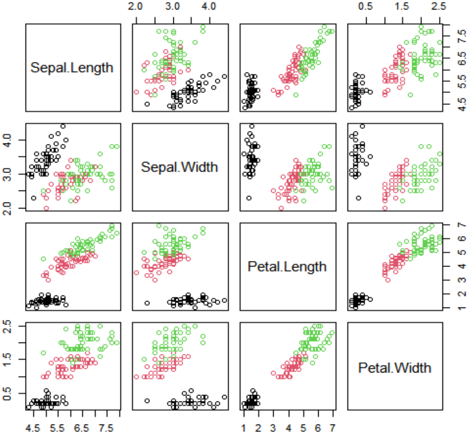


Various plotting options

- Basic *R* plot
- *lattice* library
- *ggplot* library
- 3D ploing options

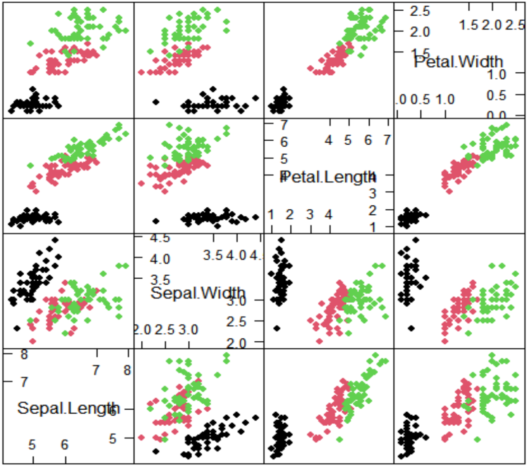
Basic *R* plot with colors for multivariate data

pairs (iris [, 1:4], col = iris$Species)



Lattice library to visualize multivariate data

```
library (lattice)
splom (  iris[ , 1:4], col = iris$Species, pch = 16)
```

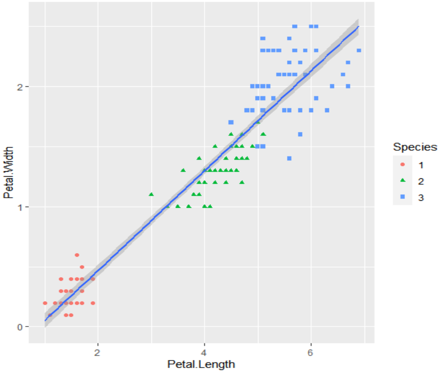


Scatter Plot Matrix

ggplot library to visualize multivariate data

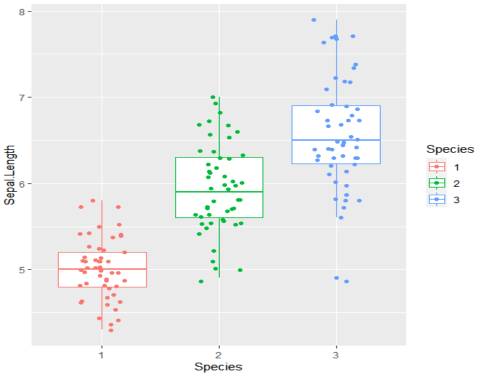
Plot petal length and petal width grouped by species with a linear trend line

```
library (ggplot2)
ggplot (data = iris) + aes(x = Petal.Length, y = Petal.Width) +
geom_point(aes(color = Species, shape = Species)) + geom_smooth(method = lm)
```



Plot the boxplot of Sepal length grouped by species and add the corresponding measurements

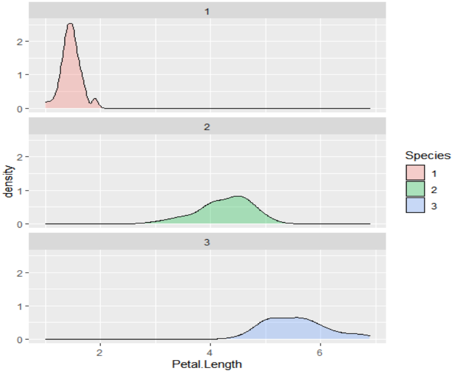
```
library (ggplot2)
ggplot(data = iris) + aes (x = Species, y = Sepal.Length, color = Species) +
geom_boxplot() + geom_jitter(position = position_jitter(0.2))
```



ggplot library to visualize multivariate data

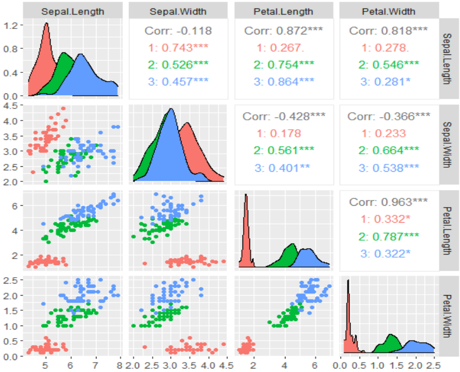
Visualize the density plot of Petal length of different species

```
library (ggplot2)
ggplot(data = iris) + aes (x = Petal.Length, fill = Species) +
geom_density(alpha = 0.3) + facet_wrap(   Species, nrow = 3)
```



Visualize correlation plot using *ggplot*

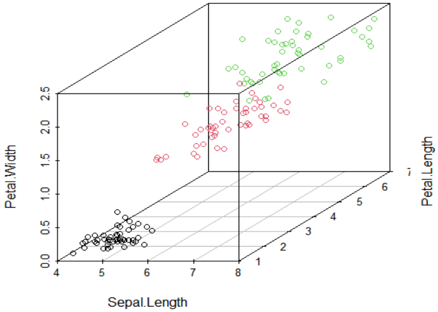
```
library (ggplot2)
library (GGally)
ggpairs (data = iris, columns = 1:4, mapping = aes (color = Species))
```



3D Plots

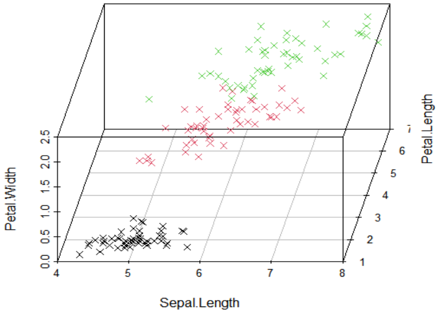
Visualize sepal length, petal length, and petal width of iris data grouped by species

```
library (scatterplot3d)
scatterplot3d (iris[ , c(1, 3, 4)], color = as.numeric(iris$Species))
```



Change the angle between X axis and y axis of the previous 3D plot to 80 degrees

```
library (scatterplot3d)
scatterplot3d(iris[, c(1, 3, 4)], color = as.numeric(iris$Species),pch = 4, angle =
80)
```



End Note: Normality Trap

- Not all data follows a Normal Distribution.
- Data with outliers or skewness may not be Normally distributed.
- Large samples will be closer to a Normal distribution than small samples.
- Real-life data is almost NEVER EXACTLY NORMAL.

References

1. Saul Stahl (2006), “The evolution of normal distribution”, Mathematics Magazine, vol. 79, no. 2, 96 - 113.
2. Kundu, Debasis. "A Journey Beyond Normality." (2014).