Introduction to R Software

Swayam Prabha

Lecture 37

Boxplots, Skewness and Kurtosis

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Summary of observations

In R, quartiles, minimum and maximum values can be easily obtained by the `summary` command

```
summary(x)  x: data vector
```

It gives information on

- **minimum,**
- **maximum**
- **first quartile**
- **second quartile (median) and**
- **third quartile.**

Summary of observations

Example:

```
> marks <- c(c(56, 59, 42, 68, 89, 29, 51, 82,
   63, 86, 34, 96, 41, 75, 77)

> summary(marks)
  Min. 1st Qu.  Median    Mean   3rd Qu.   Max.
  29.0  46.5     63.0     63.2    79.5     96.0
```

Summary of observations

Example:

R Console

```
> summary(marks)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   29.0    46.5    63.0    63.2    79.5    96.0
>
```

Summary of observations

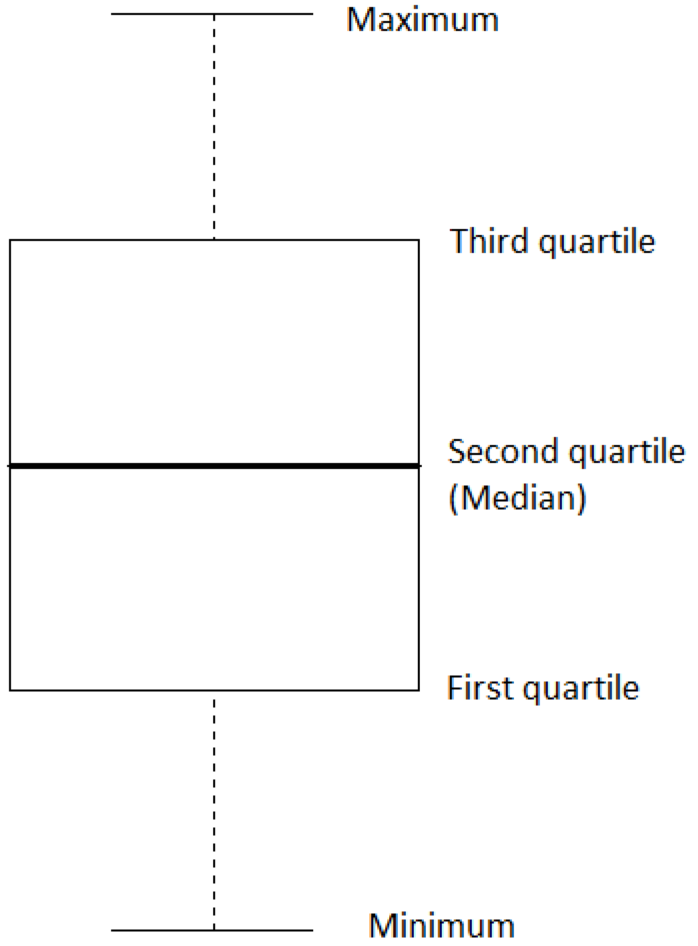
Example:

```
> marks <- c(c(56, 59, 42, 68, 89, 29, 51, 82,
   63, 86, 34, 96, 41, 75, 77)

> marks1 <- c(986, 795, 77, 56, 509, 613, 186,
   34, 41, 89, 29, 51, 420, 628, 812)

> summary(marks)
  Min. 1st Qu.  Median   Mean   3rd Qu.  Max.
  29.0    46.5   63.0    63.2    79.5    96.0

> summary(marks1)
  Min. 1st Qu.  Median   Mean   3rd Qu.  Max.
  29.0    53.5   186.0   355.1   620.5   986.0
```

Boxplot

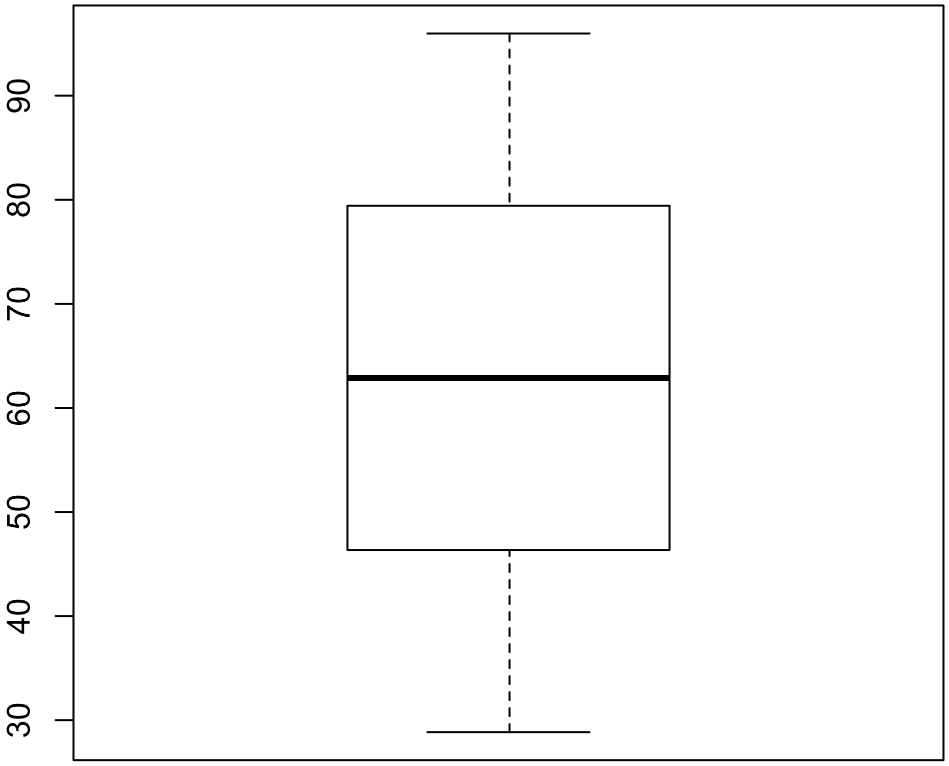
Box plot is a graph which summarizes the distribution of a variable by using its median, quartiles, minimum and maximum values.

`boxplot()` draws a box plot.



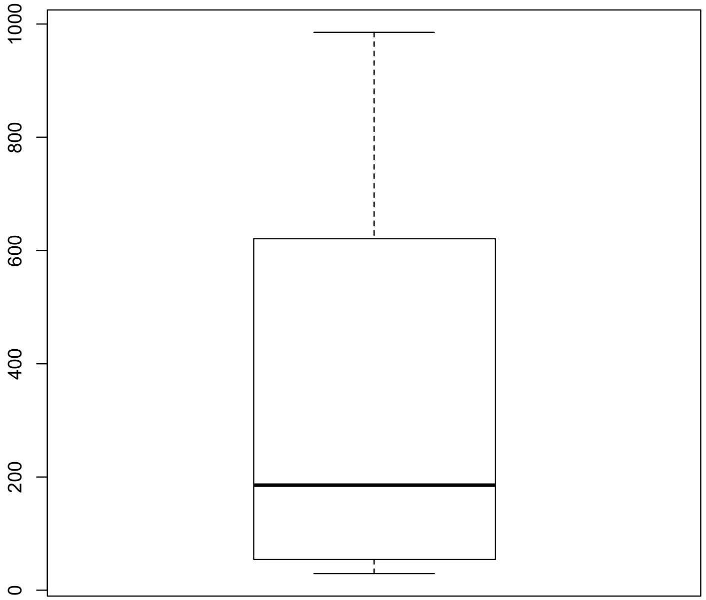
Example:

```
> marks <- c(56, 59, 42, 68, 89, 29, 51, 82,
   63, 86, 34, 96, 41, 75, 77)

> boxplot(marks)
```

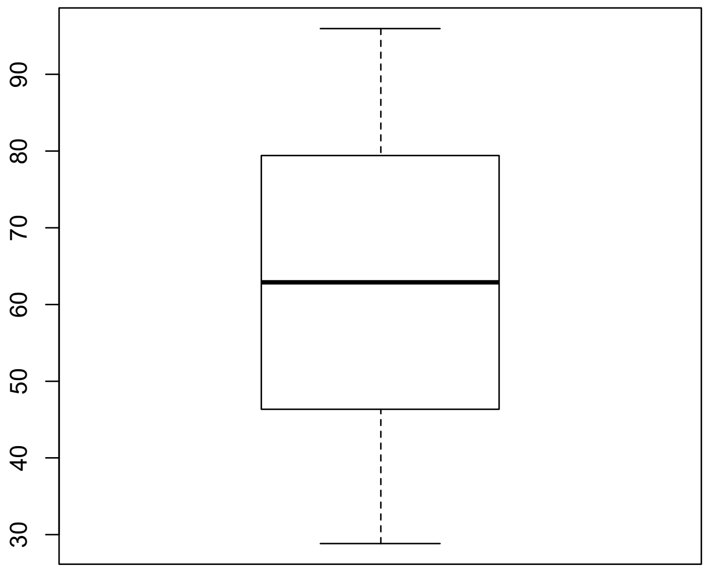


Example:

```
> marks1 <- c(986, 795, 77, 56, 509, 613, 186,
   34, 41, 89, 29, 51, 420, 628, 812)

> boxplot(marks1)
```



Boxplot(marks1)



Boxplot(marks)

Descriptive statistics:

First hand tools which gives first hand information.

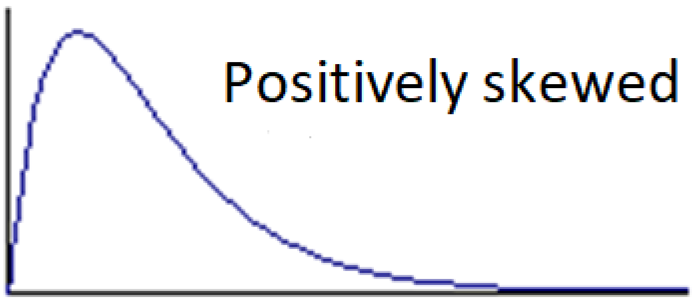
- **Structure and shape of data tendency (symmetricity, skewness, kurtosis etc.)**

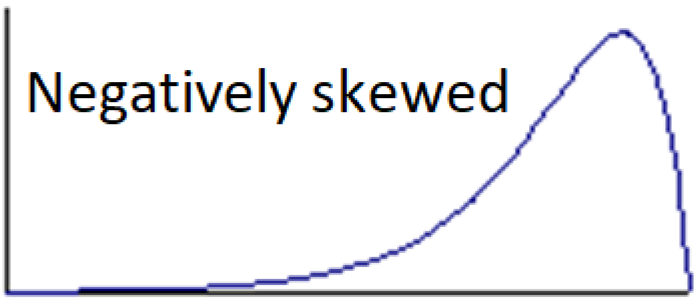
Skewness

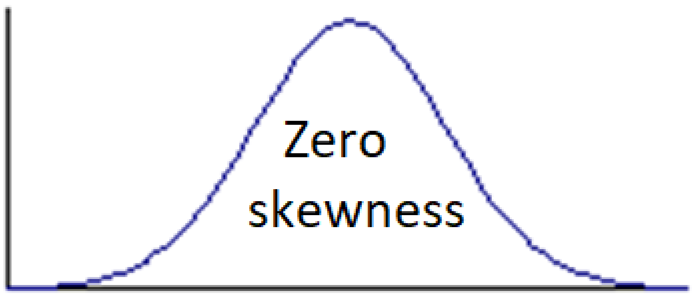
Frequency distribution for which the curve has longer tail towards the

- **right hand side is said to be positively skewed.**
- **left hand side is said to be negatively skewed.**

A symmetric curve has no or zero skewness.







Coefficient of Skewness

Sample based coefficients of skewness are

$$\beta_1 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^3}$$

Informs magnitude only.

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}}$$

Informs magnitude as well as direction.

Coefficient of Skewness

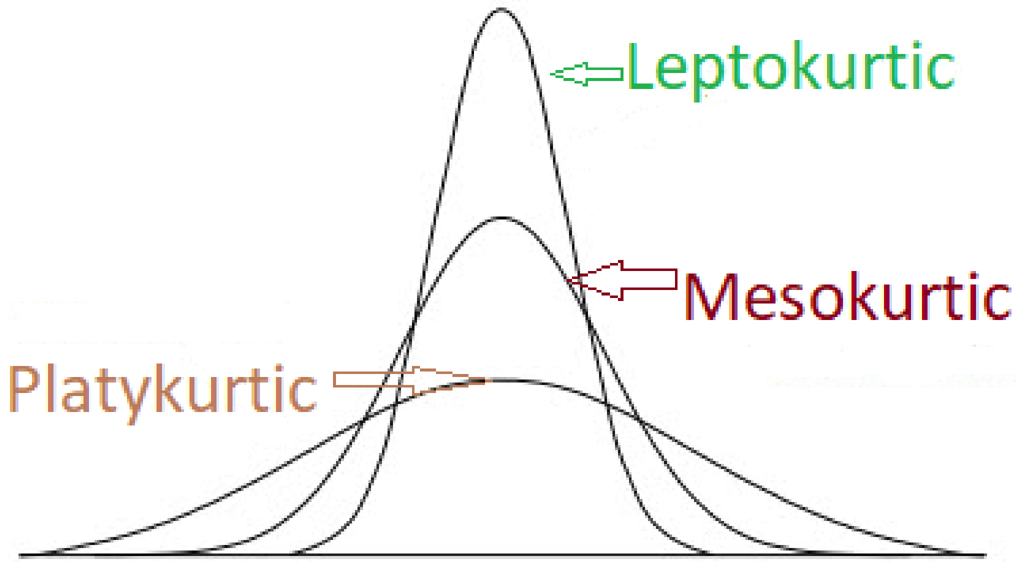
Interpretations:

Same interpretations are considered for sample based coefficients of skewness.

- **If γ_1 = 0, it means the distribution is symmetric.**
- **If γ_1 > 0, it means the distribution is positively skewed.**
- **If γ_1 < 0, it means the distribution is negatively skewed.**

Kurtosis

Observe the following curve. The three curves are representing three frequency distributions.



Kurtosis describes the peakedness ar flatness of a frequency curve.

Kurtosis

Shape of the hump (middle part of the curve or frequency distribution) of the normal distribution has been accepted as a standard.

Kurtosis examines the hump or flatness of the given frequency curve or distribution with respect to the hump or flatness of the normal distribution.

Kurtosis

Curves with hump like of normal distribution curve are called <u>mesokurtic</u>.

Curves with greater peakedness (or less flatness) than of normal distribution curve are called <u>leptokurtic</u>.

Curves with less peakedness (or grater flatness) than of normal distribution curve are called <u>platykurtic</u>.

Kurtosis

Measures the peakedness of the frequency curve.

Coefficient of kurtosis based on values $x_1, x_2, \ldots, x_n$.

$$\beta_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}\ ,\ -3 < \gamma_2 < 3$$

$$\gamma_2 = \beta_2 - 3 \qquad -\infty < \gamma_2 < \infty$$

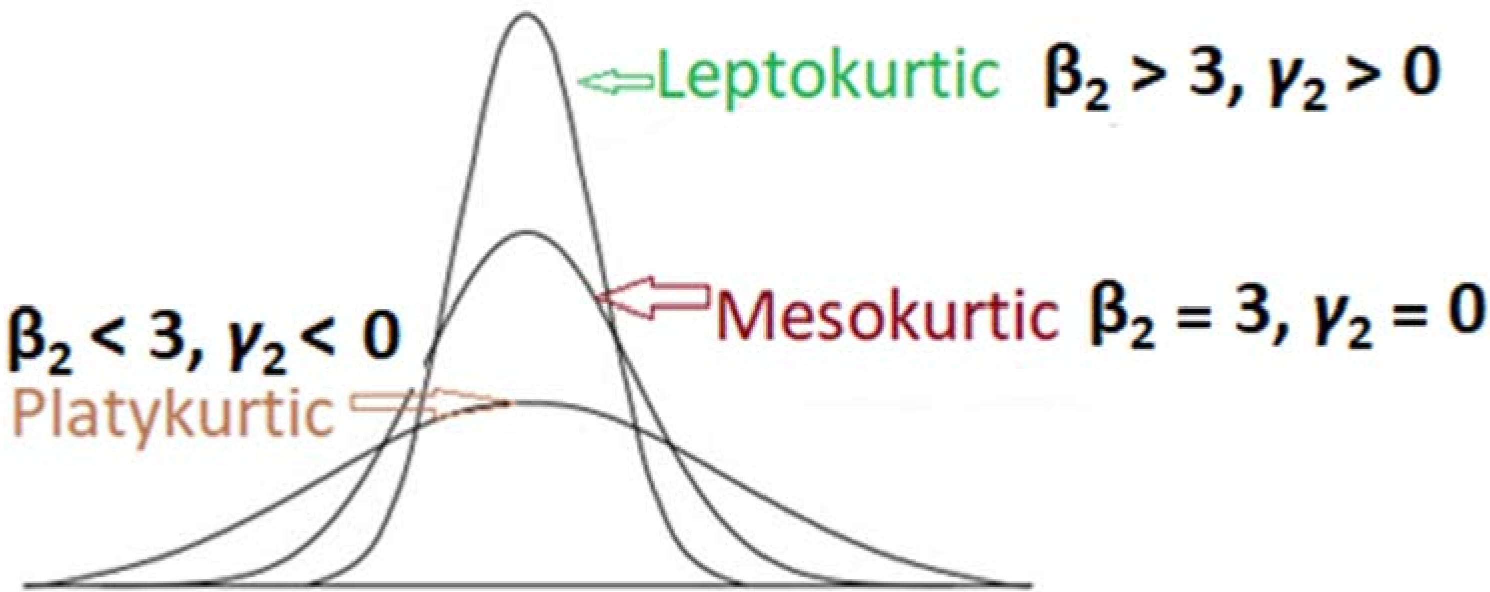
Coefficient of Kurtosis

For normal distribution, $\beta_2 = 3, \gamma_2 = 0$

For leptokurtic distribution, $\beta_2 > 3, \gamma_2 > 0$

For mesokurtic distribution, $\beta_2 = 3, \gamma_2 = 0$

For platykurtic distribution, $\beta_2 < 3, \gamma_2 < 0$



Skewness and kurtosis

First we need to install a package 'moments'

```
> install.packages("moments")

> library(moments)

skewness ()  : computes coefficient of skewness

kurtosis ()  : computes coefficient of kurtosis
```

Skewness and kurtosis

R Computes

skewness () : computes coefficient of skewness

$$\gamma_1 = \pm\sqrt{\beta_{1s}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}}$$

kurtosis () : computes coefficient of kurtosis

$$\beta_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{2}} \ , \ -3 < \gamma_2 < 3$$

Skewness and kurtosis

Example

```
> marks <- c(56, 59, 42, 68, 89, 29, 51, 82,
63, 86, 34, 96, 41, 75, 77)

> skewness(marks)
[1] -0.09869395

> kurtosis(marks)
[1] 1.830791
```