Descriptive Statistics With R Software

Moments

::

Skewness and Kurtosis

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Skewness

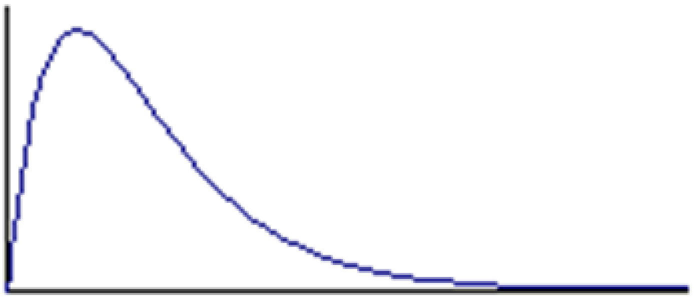
Literal meaning of skewness: Lack of symmetry

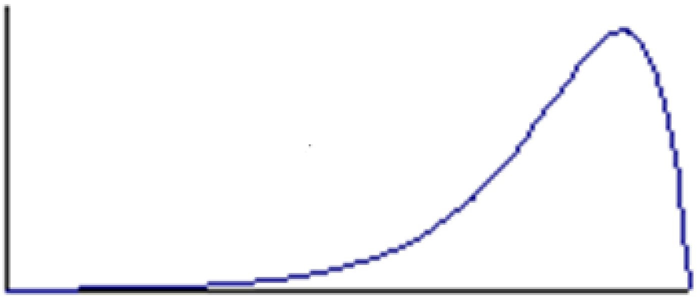
Skewness gives an idea about the

- **shape of curve obtained by frequency distribution (frequency curve) of data.**
- **nature and concentration of observations towards higher/lower values of variables.**

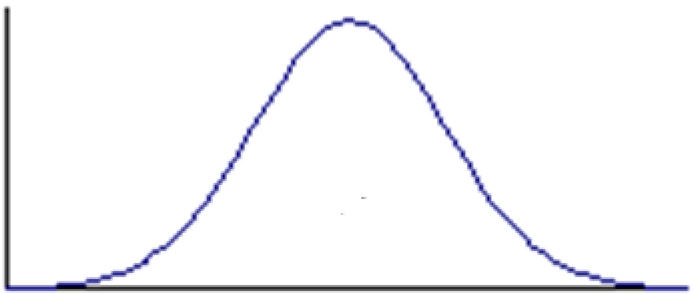
Skewness

A distribution is said to be ‘skewed‘ if the frequency curve of the distribution is not bell shaped curve and it is stretched more to one side than to the other.





A symmetric curve has no or zero skewness.

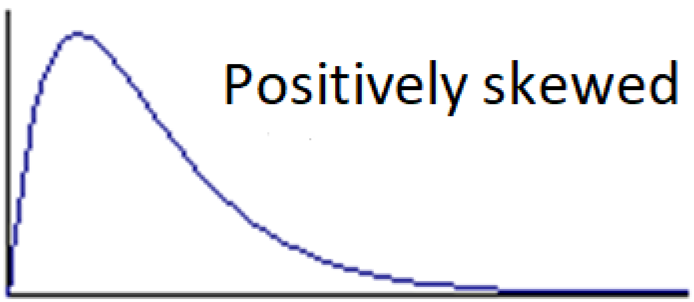


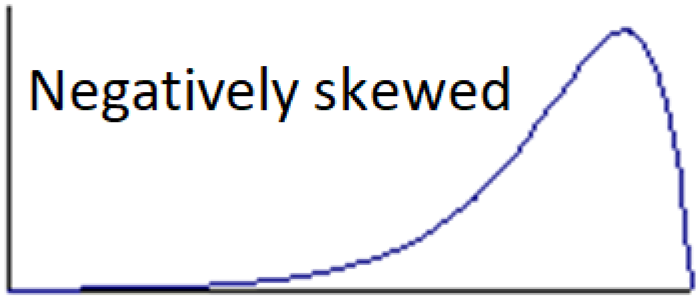
Skewness

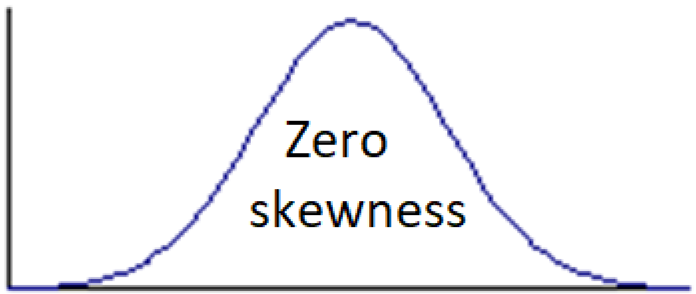
Frequency distribution for which the curve has longer tail towards the

- 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

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

where μ_2 and μ_3 are the second and third central moments respectively.

Another coefficient of skewness is

$$\gamma_1 = \pm\sqrt{\beta_1}$$

β_1 **measures the magnitude only.**

γ_1 **gives information on magnitude as well as signs as positive (+) or negative (-).**

Coefficient of Skewness

Sample based coefficients of skewness are

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

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

Coefficient of Skewness

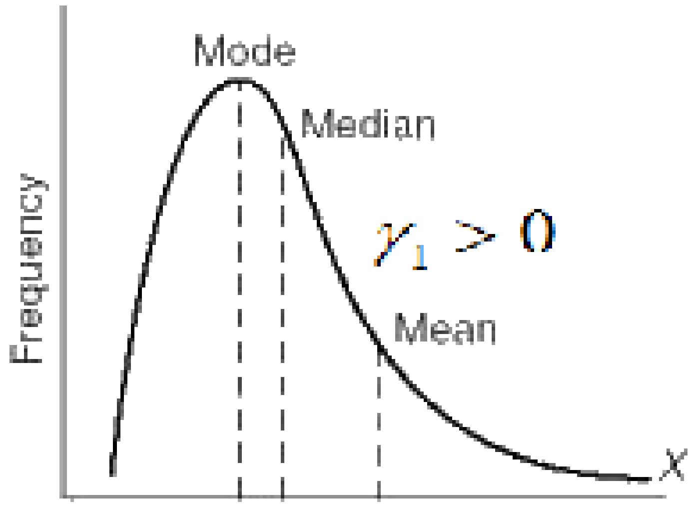
Interpretations:

- **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.**

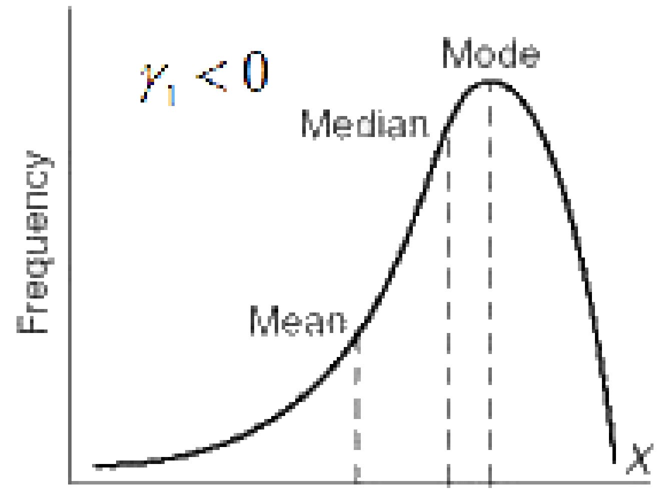
Same interpretations are considered for sample based coefficients of skewness.

- **If γ_{1s} = 0, it means the distribution is symmetric.**
- **If γ_{1s} > 0, it means the distribution is positively skewed.**
- **If γ_{1s} < 0, it means the distribution is negatively skewed.**

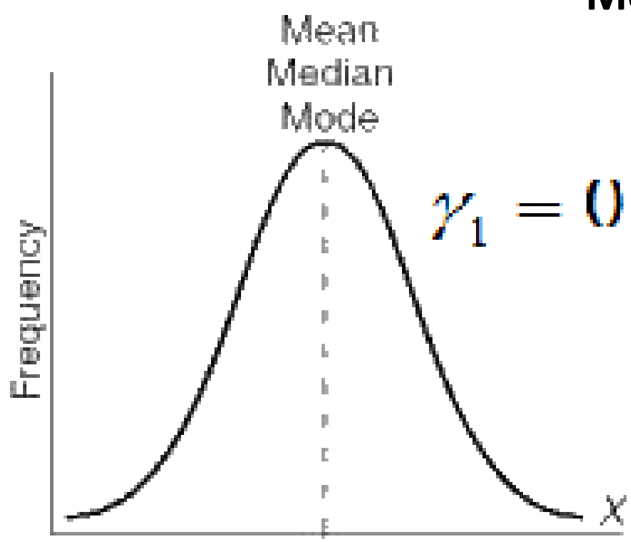
Coefficient of Skewness



Positively skewed distribution
Mode < Median < Mean



Negatively skewed distribution
Mode > Median > Mean



Symmetric distribution
Mode = Median = Mean

Coefficient of Skewness

Another coefficient of skewness is defined as

$$S_{sk} = \frac{\overline{X} - \overline{X}_{mode}}{\sigma_x} \;;\; -3 \le S_{sk} \le 3$$

$$S_{sk} = \frac{3(\overline{X} - \overline{X}_{median})}{\sigma_x} \;;\; -3 \le S_{sk} \le 3$$

- $S_{sk} > 0$ **for positively skewed distribution**
- $S_{sk} < 0$ **for negatively skewed distribution**
- $S_{sk} = 0$ **for symmetric distribution**

Coefficient of Skewness

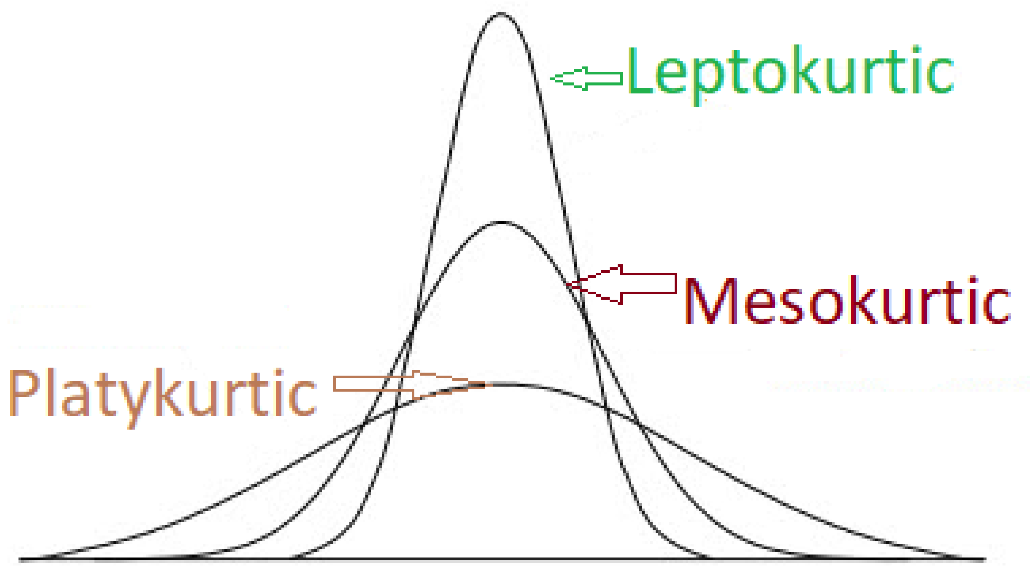
Other variants of coefficeient of skewness based on quartiles (Q_1, Q_2, Q_3, Q_4) and percentiles (P_{10}, P_{50}, P_{90}) are

$$S_{qsk} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \ ; \ -1 \le S_{qsk} \le 1$$

$$S_{psk} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} \ ; \ -1 \le S_{psk} \le 1$$

Kurtosis

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



Kurtosis describes the peakedness ar flatness of a frequency curve.

Normal Distribution

Normal Distribution : $N(\mu,\sigma^2)$

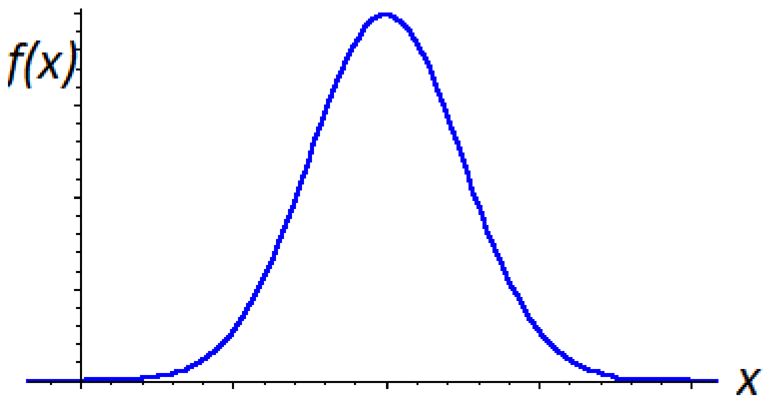
Probability distribution function

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); -\infty < x < \infty$$

μ,σ^2 : Parameters

μ : Mean, σ^2 : Variance

Bell Shaped Curve, Symmetric around mean, Skewness=0, Kurtosis=0



12

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>.

Coefficient of Kurtosis

Karl Pearson‘s coefficient of kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where μ_2 and μ_4 are the second and fourth central moments respectively.

$$\gamma_2 = \beta_2 - 3$$

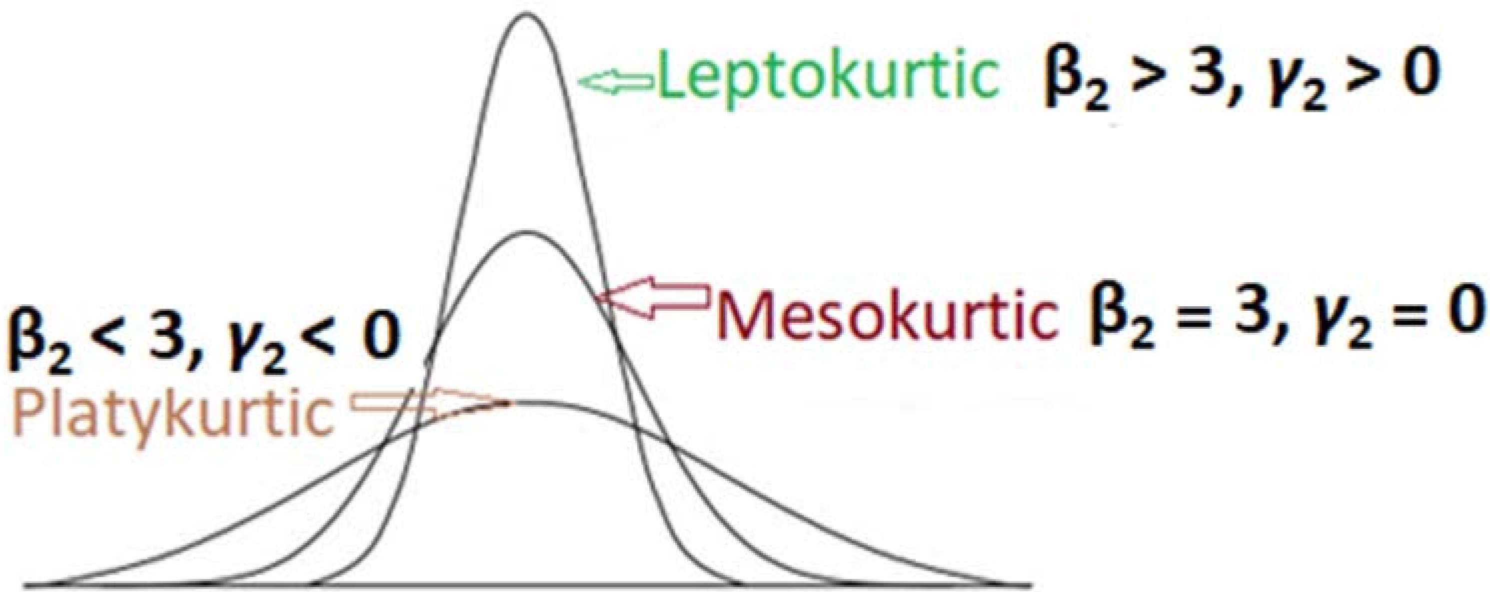
Coefficient of Kurtosis

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, $\beta_2 < 3$, $\gamma_2 < 0$



Coefficient of Kurtosis

Few properties

- $\beta_2 \geq 1$
- $\beta_2 > \beta_1$
- $\beta_2 \geq \beta_1 + 1$

Coefficient of Kurtosis

Sample based coefficients of kurtosis are

$$\beta_{2s} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}$$

$$\gamma_{2s} = \beta_{2s} - 3$$

For leptokurtic distribution, $\beta_{2s} > 3$, $\gamma_{2s} > 0$

For mesokurtic distribution, $\beta_{2s} = 3$, $\gamma_{2s} = 0$

For platykurtic distribution, $\beta_{2s} < 3$, $\gamma_{2s} < 0$

Skewness and Kurtosis

R Commands:

First we need to install a package 'moments'

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

Sample based coefficient of skewness

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

```
skewness(x, na.rm = FALSE)
```

Sample based coefficient of kurtosis

$$\gamma_{2s} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{2}} - 3$$

```
kurtosis(x, na.rm = FALSE)
```

`x` **Numeric vector, matrix or data frame.**

`na.rm` **logical TRUE if missing values need to be removed**

Skewness and Kurtosis

R Commands:
When data is missing and data vector is `x.na`

Sample based coefficient of skewness

```
skewness(xna, na.rm = TRUE)
```

Sample based coefficient of kurtosis

```
kurtosis(x.na, na.rm = TRUE)
```

`x.na` **Numeric vector, matrix or data frame containing NA values.**
`na.rm` **logical TRUE if missing values need to be removed**

Skewness and Kurtosis

Example:

Following are the time taken (in seconds) by 20 participants in a race: 32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> skewness(time)

[1] 0.05759762


> kurtosis(time)

[1] 1.701762
```

Skewness and Kurtosis

Example:

R Console

```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> skewness(time)
[1] 0.05759762
> kurtosis(time)
[1] 1.701762
>
```

Skewness and Kurtosis

Example: **Handling missing values**

Suppose two data points are missing in the earlier example where the time taken (in seconds) by 20 participants in a race. They are recorded as NA

NA, NA, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time.na = c(NA, NA, 45, 83, 74, 55, 68, 38,
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> skewness(time.na, na.rm = TRUE)
[1] -0.0614137

> kurtosis(time.na, na.rm = TRUE)
[1] 1.810021
```

Skewness and Kurtosis

Example: Handling missing values

R Console

```
> time.na
 [1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> skewness(time.na, na.rm = TRUE)
[1] -0.0614137
> kurtosis(time.na, na.rm = TRUE)
[1] 1.810021
```