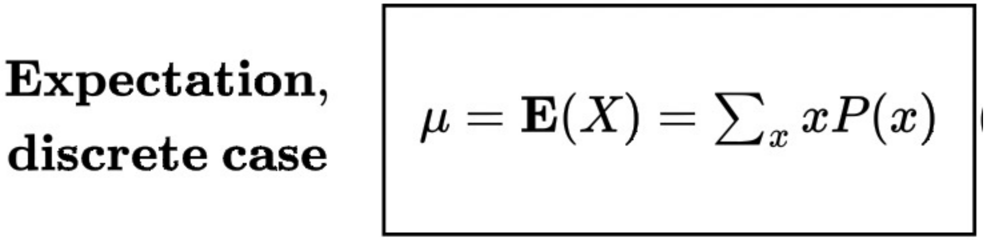
# Statistics week 4

# Expected Value

- Expectation or expected value of a random variable X is its mean, the average value.
- We know that X can take different values with different probabilities. For this reason, its average value is not just the average of all its values. Rather, it is a weighted average.



**Expectation, discrete case**

$$\boxed{\mu = \mathbf{E}(X) = \sum_x x P(x)}$$

# Example-Expected value

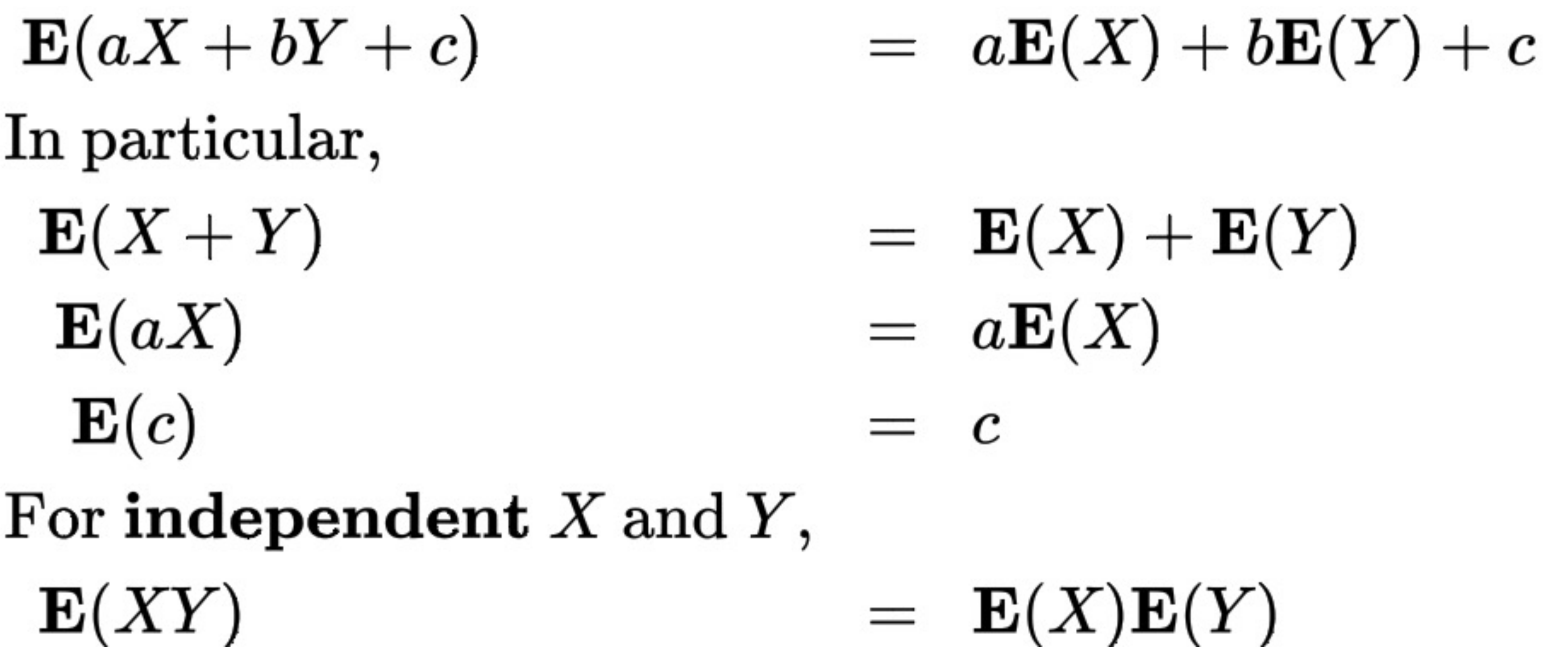
- If you roll a 6 sided die , each side has a probability of 1/6 for landing.
- The Expected value is given as:

$$\left(\frac{1}{6}\times 1\right)+\left(\frac{1}{6}\times 2\right)+\left(\frac{1}{6}\times 3\right)$$
$$+\left(\frac{1}{6}\times 4\right)+\left(\frac{1}{6}\times 5\right)+\left(\frac{1}{6}\times 6\right)=3.5$$

# Expectation of function

$$\mathbf{E}\{g(X)\} = \sum_x g(x)P(x)$$

# Properties of Expectation



$$\mathbf{E}(aX + bY + c) = a\mathbf{E}(X) + b\mathbf{E}(Y) + c$$

In particular,

$$\begin{aligned} \mathbf{E}(X + Y) &= \mathbf{E}(X) + \mathbf{E}(Y) \\ \mathbf{E}(aX) &= a\mathbf{E}(X) \\ \mathbf{E}(c) &= c \end{aligned}$$

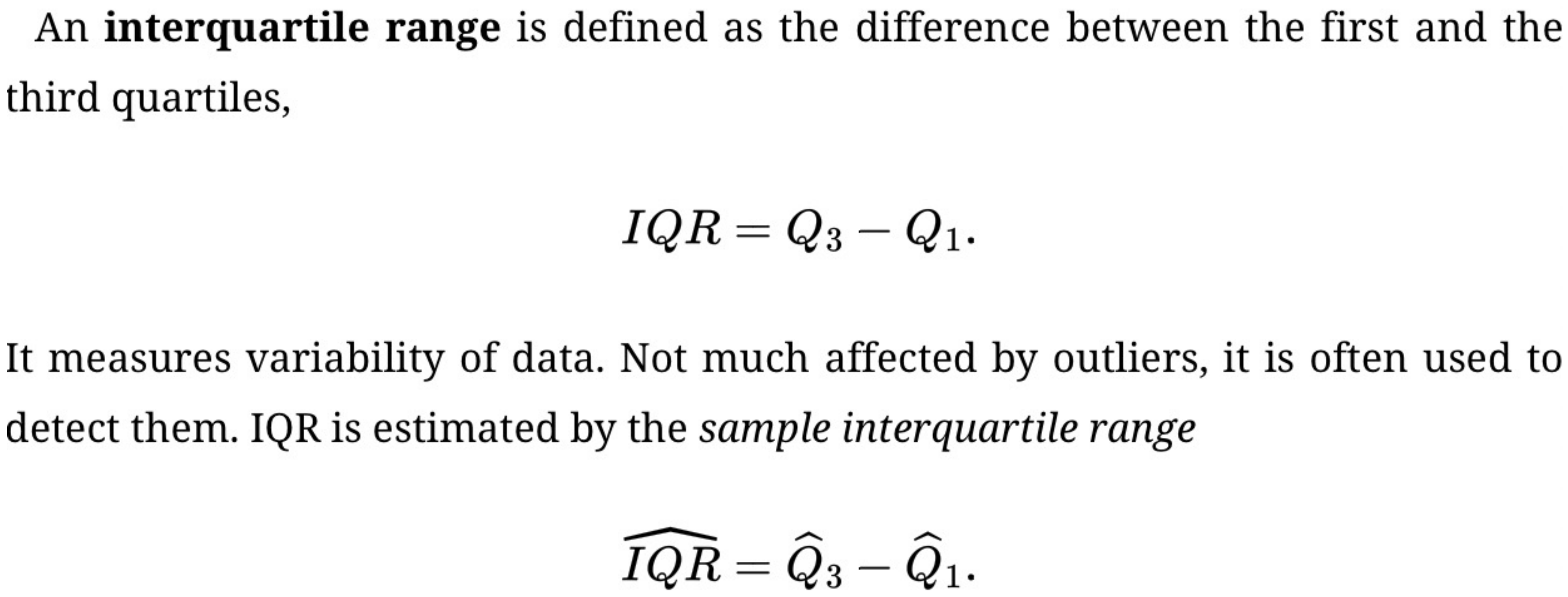
For **independent** $X$ and $Y$,

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$$

# Interquartile Range

- Outliers are values that are abnormally different and away from the observed values.
- An outlier can be because of an error, variability in measurement or can sometimes represent a new phenomenon.
- If an extreme observation (an outlier) erroneously appears in our data set, it can rather significantly affect the values of mean and standard deviation.
- In practice, outliers may be a real problem that is hard to avoid. To detect and identify outliers, we need measures of variability that are not very sensitive to them.
- One such measure is interquartile range.

# Interquartile Range



An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1.$$

It measures variability of data. Not much affected by outliers, it is often used to detect them. IQR is estimated by the *sample interquartile range*

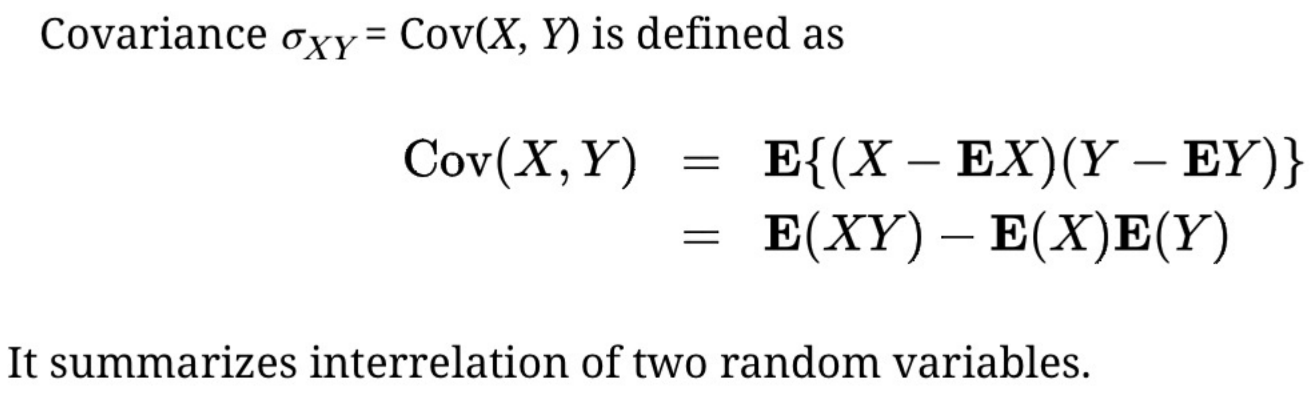
$$\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1.$$

# Interquartile Range

- A “rule of thumb” for identifying outliers is the rule of 1.5(IQR). Measure 1.5(Q^3-Q^1) down from the first quartile and up from the third quartile. All the data points observed outside of this interval are assumed suspiciously far. They are the first candidates to be handled as outliers.

# Covariance

- Covariance is a way of measuring the relationship or association of 2 random variables.

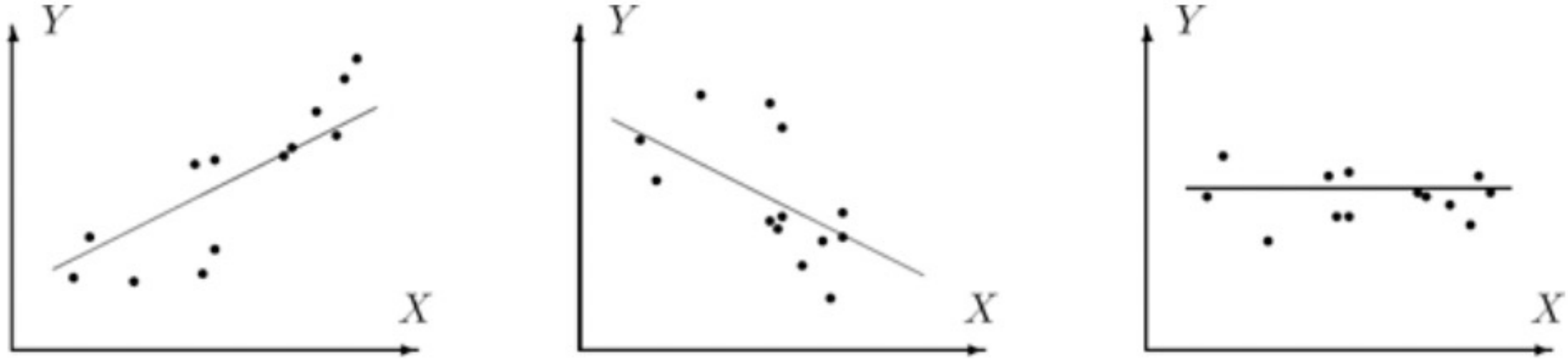


Covariance $\sigma_{XY}$ = Cov($X$, $Y$) is defined as

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbf{E}\{(X-\mathbf{E}X)(Y-\mathbf{E}Y)\} \\ &= \mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)\end{aligned}$$

It summarizes interrelation of two random variables.

# Positive, negative and zero Covariance



(a) $\mathrm{Cov}(X, Y) > 0$

(b) $\mathrm{Cov}(X, Y) < 0$

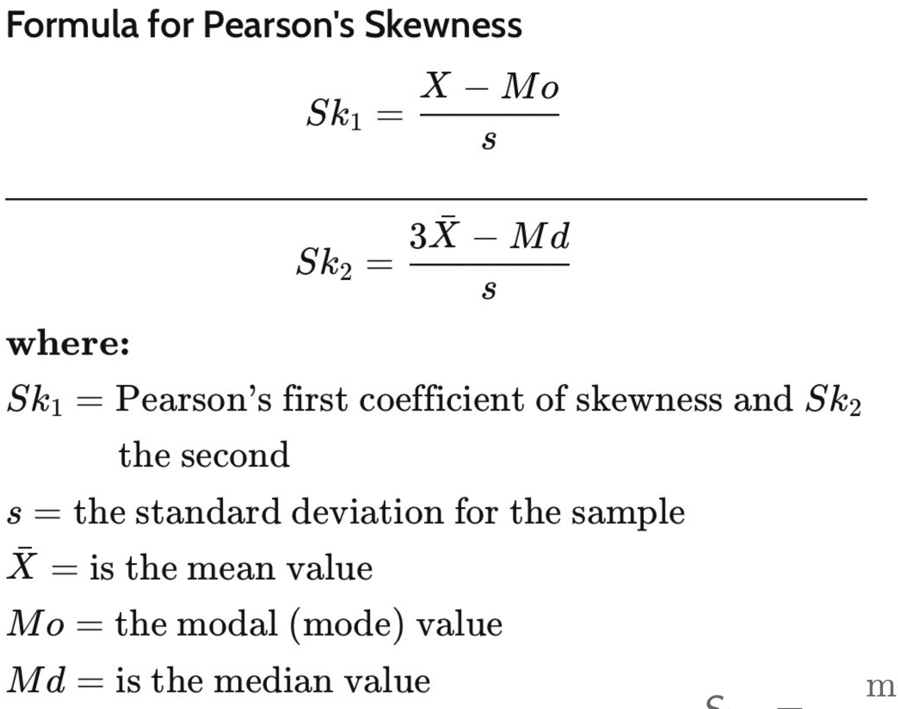
(c) $\mathrm{Cov}(X, Y) = 0$

# Skewness

- It is the degree of asymmetry in any probability distribution.
- A normal distribution, bell curve, has 0 skewness.
- Skewness shows the direction of outliers.
- It also shows how condensed is the data on any particular side.
- Comparing the mean μ and the median M, one can tell whether the distribution of X is right-skewed, left-skewed, or symmetric.

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right} - \text{skewed distribution} & \Rightarrow & M > \mu \\ \text{left} - \text{skewed distribution} & \Rightarrow & M > \mu \end{array}$$

# How to calculate Skewness



**Formula for Pearson's Skewness**

$$Sk_1 = \frac{X - Mo}{s}$$

$$Sk_2 = \frac{3\bar{X} - Md}{s}$$

**where:**

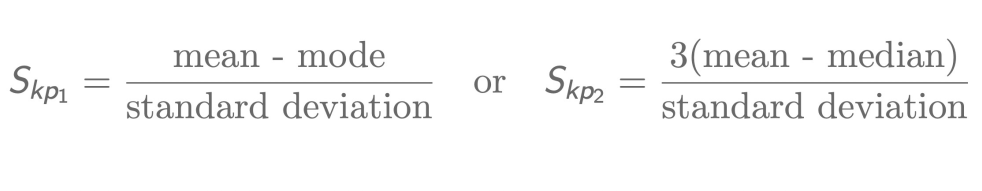
$Sk_1 =$ Pearson's first coefficient of skewness and $Sk_2$ the second

$s =$ the standard deviation for the sample

$\bar{X} =$ is the mean value

$Mo =$ the modal (mode) value

$Md =$ is the median value



$$S_{kp_1} = \frac{\text{mean - mode}}{\text{standard deviation}} \quad \text{or} \quad S_{kp_2} = \frac{\text{3(mean - median)}}{\text{standard deviation}}$$

# Example

Calculate Karl Pearson coefficient of skewness of the following data set $(S = 1.7)$.

| Value (x) | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Frequency (f) | 2 | 3 | 4 | 4 | 6 | 4 | 2 |

# Solution

$$\begin{aligned}
\text{mean} = \overline{x} &= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i}^{n} f_i} \\
&= \frac{1 \times 2 + 2 \times 3 + \ldots + 7 \times 2}{25} \\
&= \frac{104}{25} \\
&= 4.16 \\
\text{mode} &= 5 \\
S_{kp} &= \frac{\text{mean-mode}}{\text{standard deviation}} \\
&= \frac{4.16 - 5}{1.7} = -0.4941
\end{aligned}$$

Since $S_{kp} < 0$ distribution is skewed left.

# How to calculate skewness

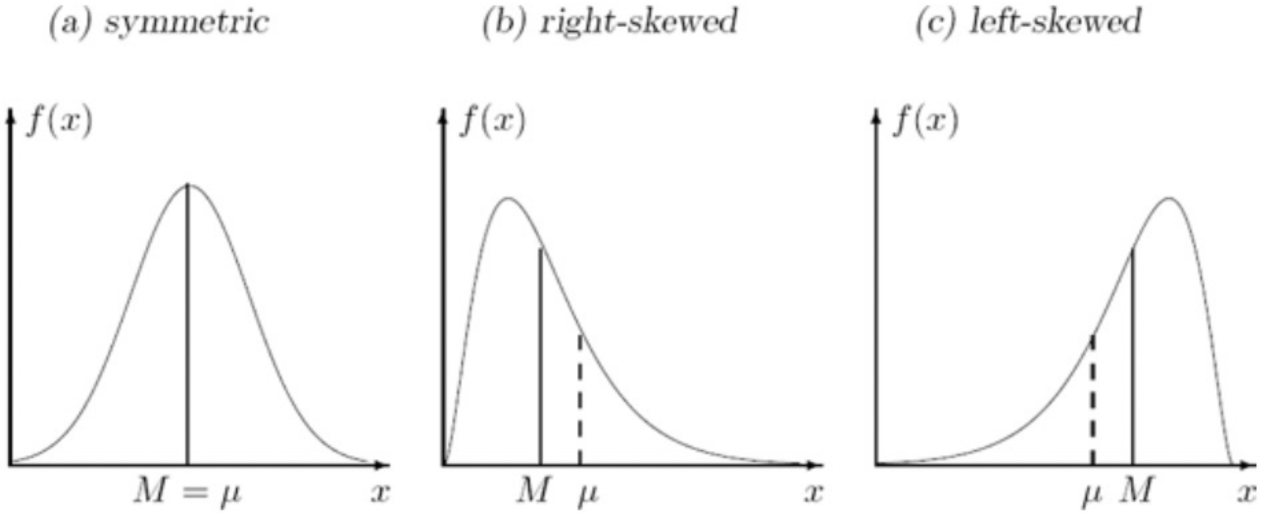
$-1 \leq Skp \leq 1$.

$Skp = 0 \Rightarrow$ distribution is symmetrical about mean.

$Skp > 0 \Rightarrow$ distribution is skewed to the right.

$Skp < 0 \Rightarrow$ distribution is skewed to the left.

# Skewness


(a) symmetric
(b) right-skewed
(c) left-skewed
$f(x)$
$f(x)$
$f(x)$
$M = \mu$
$x$
$M$
$\mu$
$x$
$\mu$
$M$
$x$

# Kurtosis

- Kurtosis is a measure what extent any distribution has outliers.
- It involves looking for extreme values in the tails.
- There are 3 types of Kurtosis:
  - Leptokurtic
  - Platykurtic
  - Mesokurtic

All three types are relative to the normal distribution.

A higher kurtosis value means that there are more outliers which fall away from the mean.

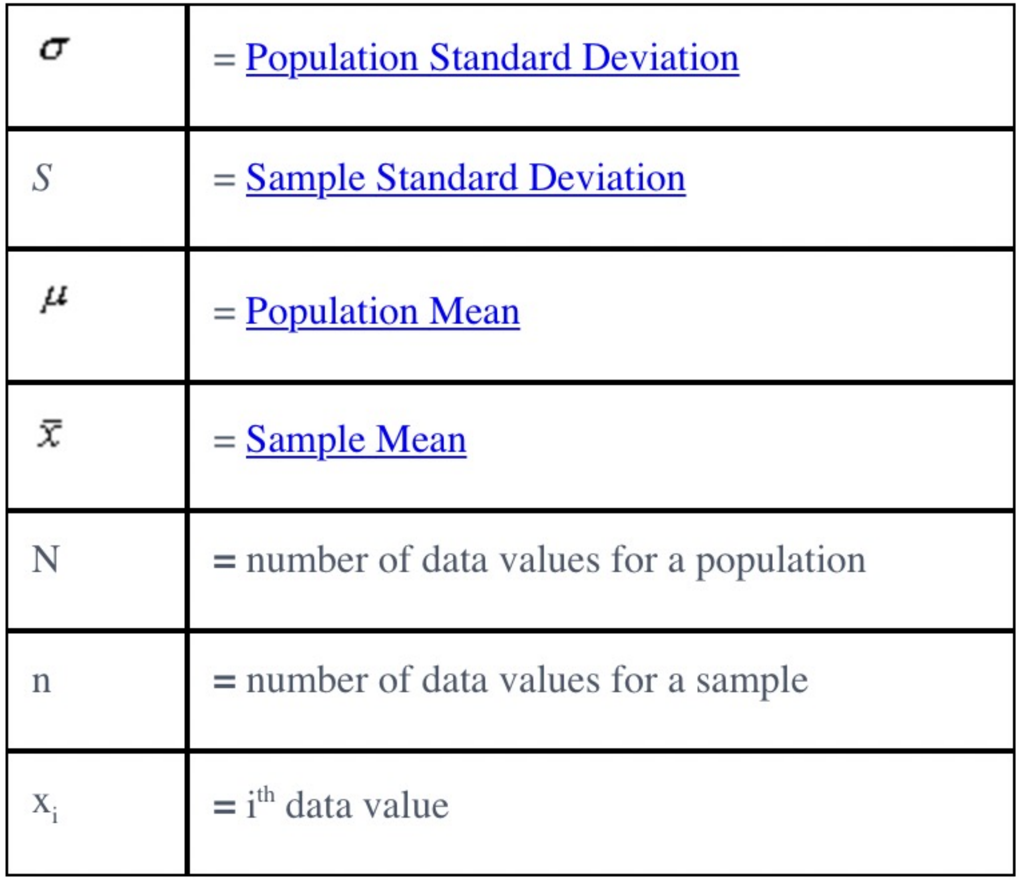
# Kurtosis

- It is also a measure of the sharpness of the peak.
- A value greater than 0 indicates a peaked distribution and a value less than 0 indicates a flat distribution.

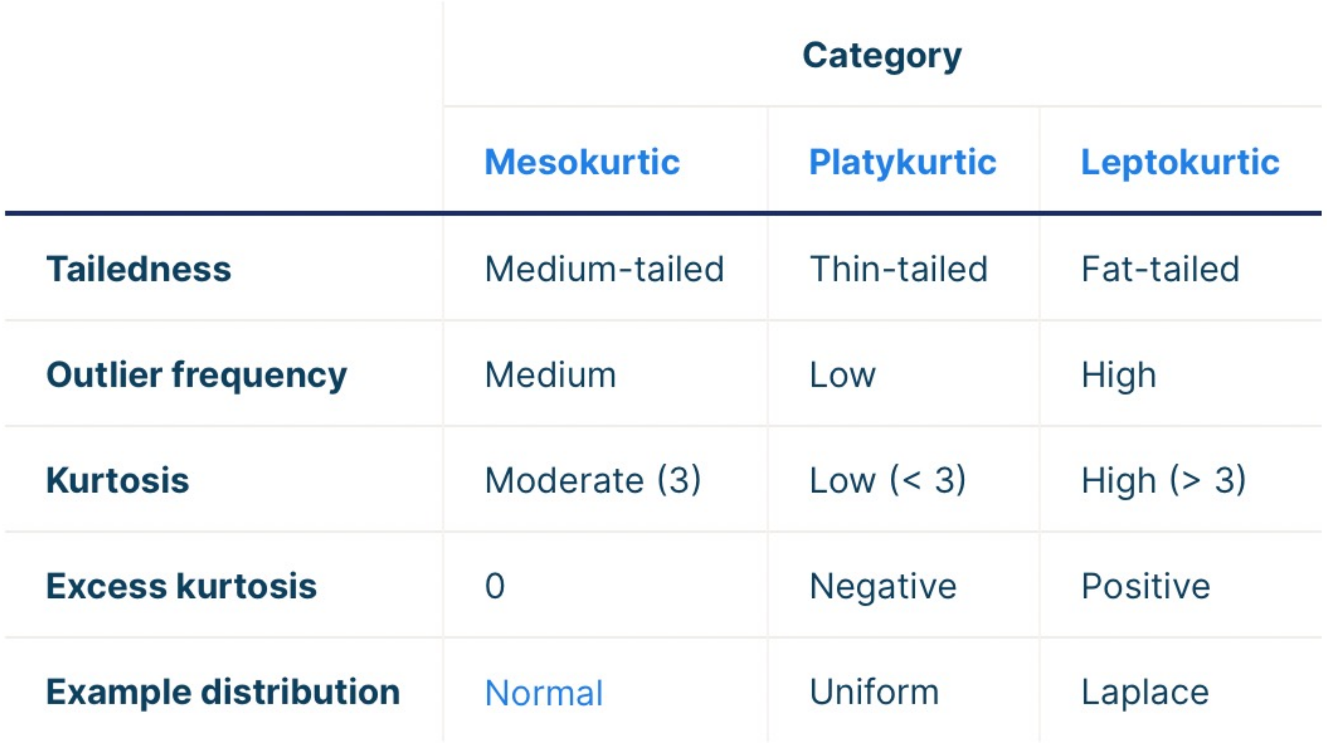
$$\gamma_2 = \left( \frac{1}{N\,\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^4 \right) - 3$$

Sample Kurtosis ($g_2$)

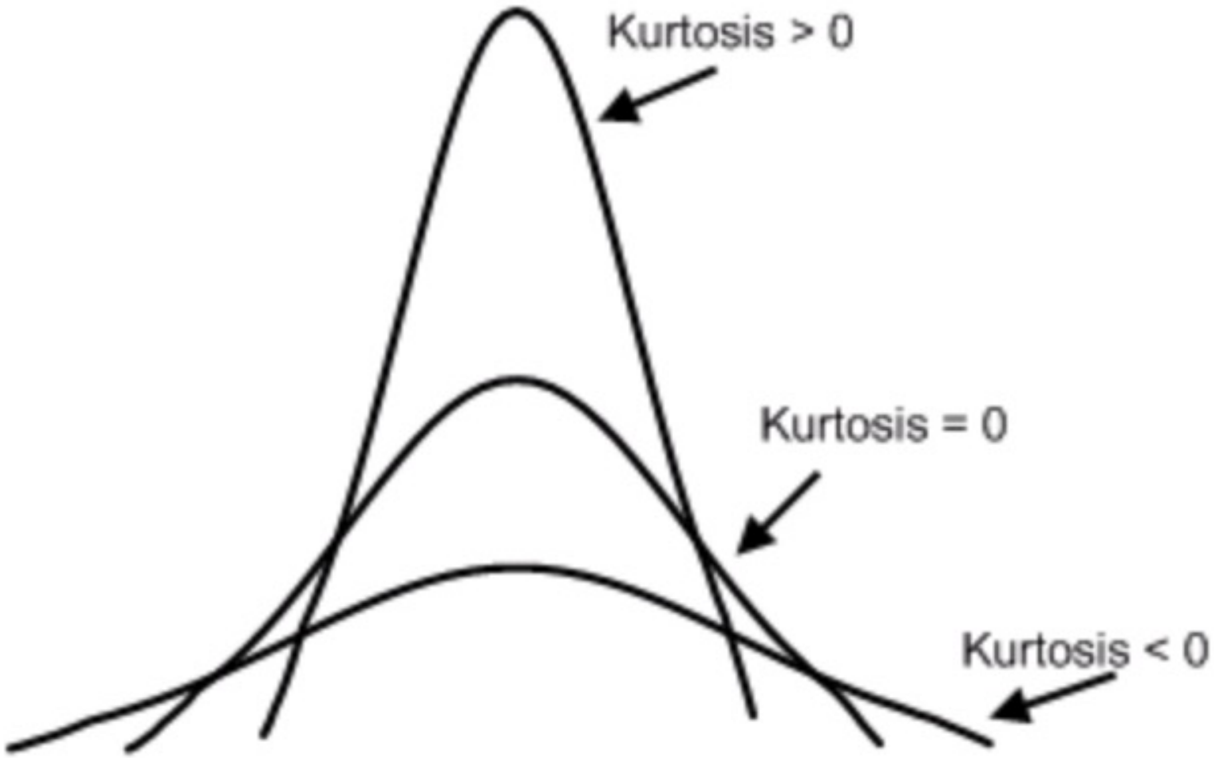
$$g_2 = \left\{ \frac{n\,(n+1)}{(n-1)(n-2)(n-3)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3\,(n-1)^2}{(n-2)(n-3)}$$



| | |
|---|---|
| $\sigma$ | = Population Standard Deviation |
| $S$ | = Sample Standard Deviation |
| $\mu$ | = Population Mean |
| $\bar{x}$ | = Sample Mean |
| N | = number of data values for a population |
| n | = number of data values for a sample |
| $x_i$ | = $i^{th}$ data value |



| | Category | | |
|---|---|---|---|
| | **Mesokurtic** | **Platykurtic** | **Leptokurtic** |
| **Tailedness** | Medium-tailed | Thin-tailed | Fat-tailed |
| **Outlier frequency** | Medium | Low | High |
| **Kurtosis** | Moderate (3) | Low (< 3) | High (> 3) |
| **Excess kurtosis** | 0 | Negative | Positive |
| **Example distribution** | Normal | Uniform | Laplace |


Kurtosis > 0
Kurtosis = 0
Kurtosis < 0

# Monte Carlo Method

- Predictive Approach
- Probability Distribution
- Repetitive simulations
- Based on computer simulations involving random numbers
- The main purpose of simulations is estimating such quantities whose direct computation is complicated, risky, consuming, expensive, or impossible.

# Monte Carlo Methods

- Monte Carlo methods are mostly used for the computation of probabilities, expected values, and other distribution characteristic.
- "A Monte Carlo simulation requires assigning multiple values to an uncertain variable to achieve multiple results and then averaging the results to obtain an estimate."
- Its applications include:
  - Physical sciences
  - Engineering
  - Climate change
  - Computational biology
  - Computer graphics
  - Artificial intelligence
  - Gaming
  - Design and visuals
  - Finance
  - Law

# Components of a Monte Carlo Simulation

- Input variables
- Output variables
- Mathematical Model

# Steps in Monte Carlo Methods

- Establish the mathematical model
- Determine the input values
- Create a sample dataset
- Set up the Monte Carlo simulation software
- Analyze the results

# Challenges of the Monte Carlo Simulations

- Choosing the right input and probability distribution
- Needs excessive computational power to run the experiment.

# Any Questions?