"Outlier Detection" - IQR

$D = [17, 17, 18, 19, 20, 22, 23, 25, 83, 64]$ (64 → outlier)

$Q_2 = 21$

$Q_1 = 18$

$Q_3 = 25$

$IQR = 7$

$$\text{Upper limit} = Q_3 + 1.5\,IQR$$
$$= 25 + 1.5 \times 7 = 35.5$$

$$\text{Lower limit} = Q_1 - 1.5\,IQR$$
$$= 18 - 1.5 \times 7 = 7.5$$

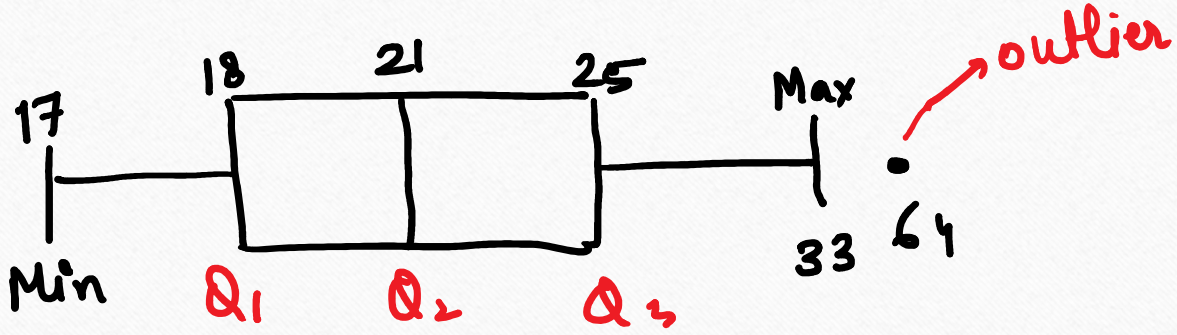
[17, 17, 18, 19, 20, 22, 23, 25, 33, 64]

$Q_2 = 21$ $Q_3 = 25$ $UL = 35.5$

$Q_1 = 18$ $IQR = 7$ $LL = 7.5$

ASSIGNMENT

" Perform this activity in Jupyter Notebook"



Quartiles: Q_1 Q_2 Q_3
25%. 50%. 75%.

Percentiles

- The kth percentile is the value that's k% of the way through your data. It's denoted by Pk
- First of all, line all your values up in ascending order.
- To find the position of the kth percentile out of n numbers, start off by calculating . $k(n/100) = Pk$
- If this gives you an integer, then percentile is halfway between the value at position and the next number along. Take the average of the numbers at these two positions to give you your percentile.
- If Pk is not an integer, then round it up. This then gives you the position of the percentile.

Checking for Outliers

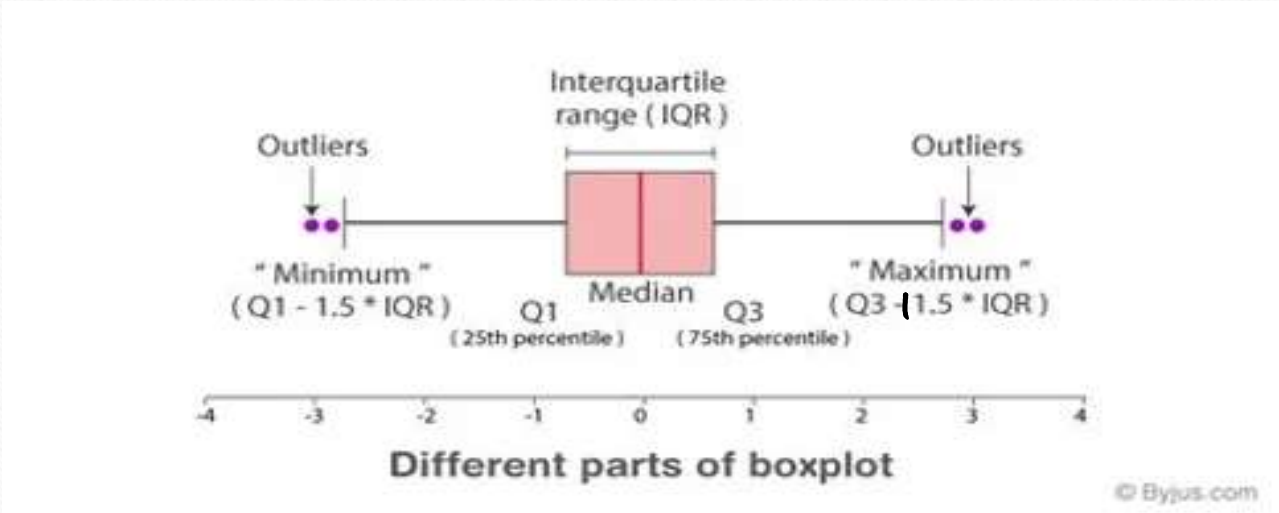
- With Range, Outlier detection isn't possible but with IQR it is possible!!

$$L = Q1 - (1.5 * IQR)$$

$$H = Q3 + (1.5 * IQR)$$

- Where L is the lower outlier
- H is the higher outlier
- Q1 and Q3 are the average values of those quartiles
- IQR is the interquartile range

Box Plot & 5 Number Summary



Variance

$Var(A) = \frac{(-5-0)^2 + (0-0)^2 + (5-0)^2}{3} = \frac{50}{3}\ km^2$

$Var(B) = \frac{(-10-0)^2 + (0-0)^2 + (10-0)^2}{3} = \frac{200}{3}\ km^2$

- The variance is a way of measuring spread, and it's the average of the distance of values from the mean squared.

population

$$VAR = \frac{1}{n}\sum_{i=0}^{n}(x - \mu)^2$$

sample

$$VAR = \frac{1}{n-1}\sum_{i=0}^{n}(x - \bar{x})^2$$

sample mean

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$v = \frac{\sum_{i=1}^{n} x_i^2}{n} - \mu^2$$

- The problem with the variance is that it can be quite difficult to think about spread in terms of distances squared.

$\sqrt{16.66 km^2} = 4.$ μ: population mean

Standard Deviation

- Standard Deviation the square root of the variance.
- It is more intuitive than Variance.
- Low standard deviation means points are closer to mean and high standard deviation means points are far from mean.

$$\text{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x-\mu)^2}$$

$$\text{S.D.} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x-\bar{x})^2}$$

MSD : [40, 45, 35, 41, 39, 40, 40] in 4overs

VK : [0, 50, 10, 60, 5, 2, 1] in "

4overs?

MSD is chosen to bat because less variation
↓
low std dev.

Measure of central tendency

[10, 10, 10, 10] ⇒ 10

−2 ↓ +2

[12, 12, 12, 12] ⇒ 12

[8, 8, 8, 8] ⇒ 8

[20, 20, 20, 20] ⇒ 20

×2 ÷2

[5, 5, 5, 5] = 5

±0

Measure of spread

−2 +1

[1, 2, 3, 4, 5] ⇒ D

−1 +2

D+1 ⇒ [2, 3, 4, 5, 6]

−2 −1 +1 +2

D−1 ⇒ [0, 1, 2, 3, 4] ⇒ No change in spread

−4 +4

D×2 = [2, 4, 6, 8, 10]

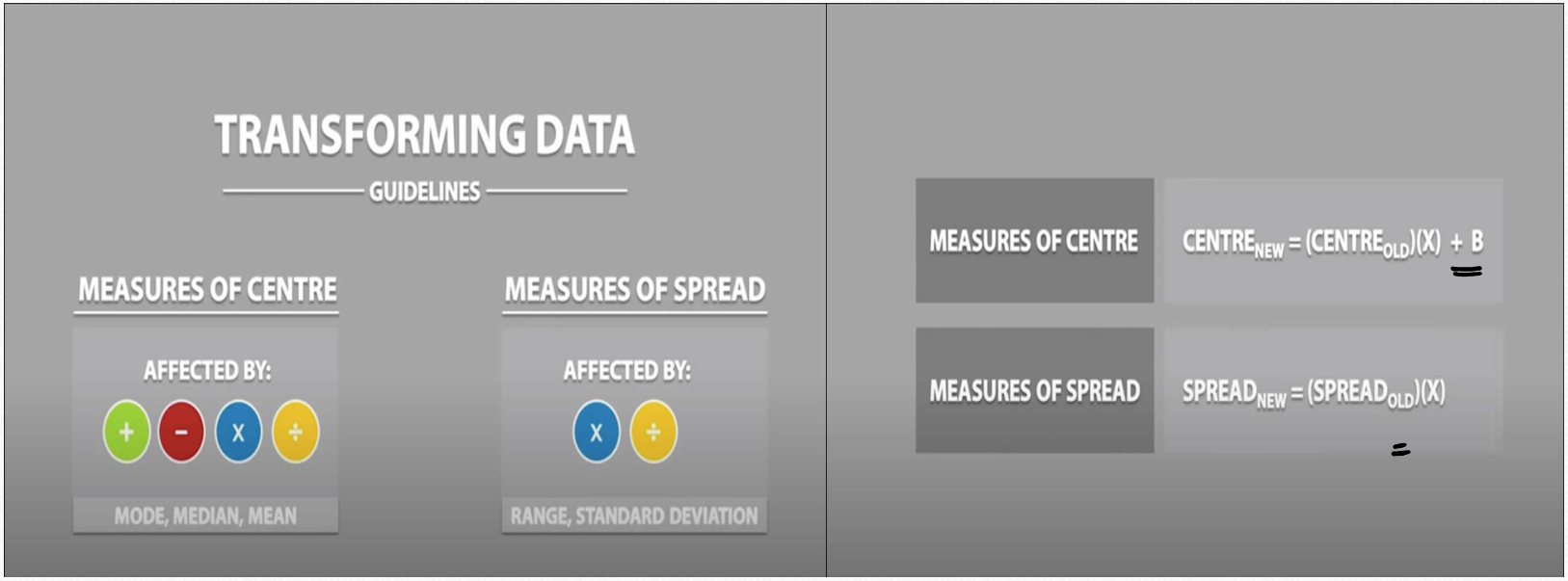
−2 +2

−1 +0.5

D÷2 = [0.5, 1, 1.5, 2, 2.5]

−0.5 +1

Transformations



Given:

$\bar{x} = 135.6$

$s = 31.75$

Example

$\bar{x}_{new} = 135.6 \times 2.5 + 750$

$S = 31.75 \times 2.5$

SUPPOSE THAT IN ORDER TO STAY HYDRATED, THESE STUDENTS DRINK 2.5mL OF WATER FOR EVERY POUND THEY WEIGH; PLUS 750mL OF WATER A DAY. WHAT IS THE MEAN AND STANDARD DEVIATION FOR THE AMOUNT OF WATER CONSUMED EVERY DAY?

105 $\bar{x} = 135.6$ × 2.5 + 750

156 $\bar{x}_{NEW} = (135.6)(2.5) + 750 = 1089$

145 $s = 31.75$ × 2.5

172 $s_{NEW} = (31.75)(2.5) = 79.38$

100

Mean = 4 $\quad \frac{1}{7}\left[(1-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2\right]$

Question to Play!

- Calculate the mean and standard deviation for the following sets of numbers:
  - 1 2 3 4 5 6 7
- The mean for each of them is 10. Your job is to play like you're the coach, and work out the standard deviation for each player. Which player is the most reliable one for your team?

A
- Score 7 9 10 11 13 Frequency 1 2 4 2 1

var(A) = 2 , std dev = 1.414

B
- Score 7 8 9 10 11 12 13 Frequency 1 1 2 2 2 1 1

var(B) = 3 , " = 1.732

a) 7 9 10 11 13 $\Rightarrow x$

$f \Rightarrow$ 1 2 4 2 1

$\mu = 10$

1 2 3 4 5 6 7 8 9 10

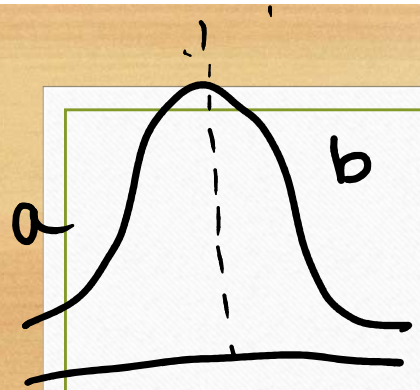
[7, 9, 9, 10, 10, 10, 10, 11, 11, 13]

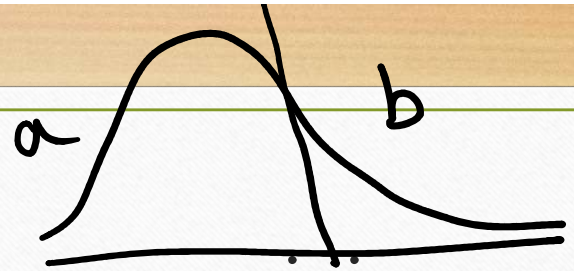
$$\text{Var} = \frac{\Sigma x^2 f}{n} - \mu^2 \Rightarrow \frac{\Sigma (x-\mu)^2}{n} \Rightarrow \frac{\Sigma (x-\mu)(x-\mu)}{n}$$

$$= \frac{7^2 \times 1 + 9^2 \times 2 + 10^2 \times 4 + 11^2 \times 2 + 13^2 \times 1}{10} - 100$$

$$= \frac{49 + 16\ 2 + 400 + 242 + 169}{10} - 100 = 102.2 - 100$$

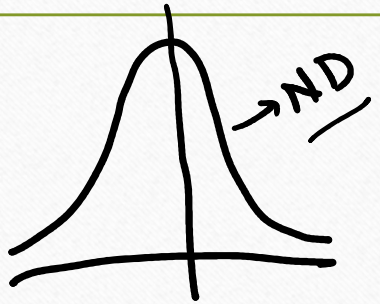
$$= 2.2$$





Measures of Symmetricity

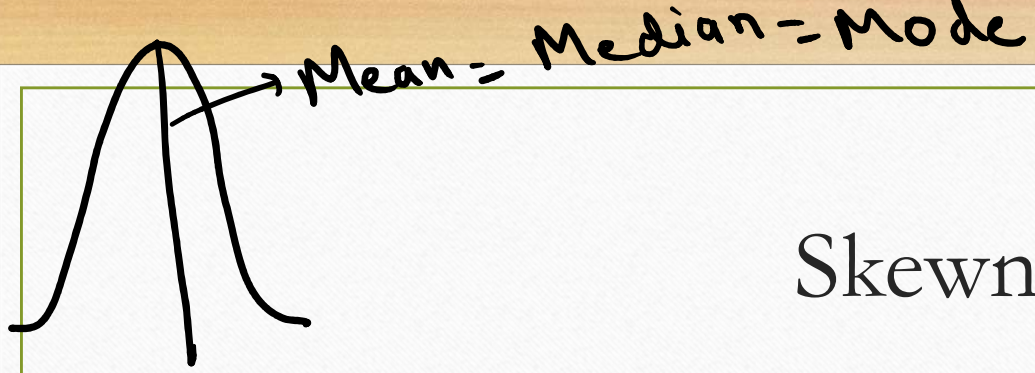
- Measures
  - Skewness
  - Kurtosis



Skewness

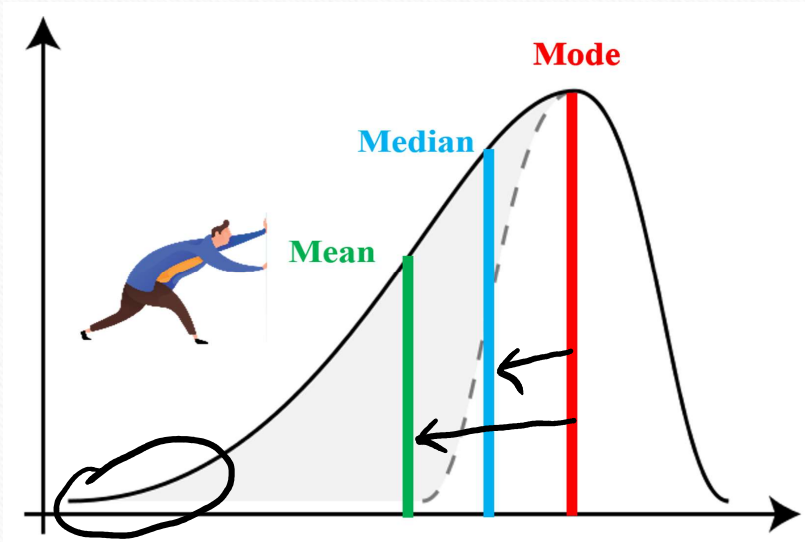
- Skewness is usually described as a measure of a dataset's symmetry or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0.
- The normal distribution has a skewness of 0.
- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
- If the skewness is less than -1 or greater than 1, the data are highly skewed

$$\gamma_1 = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^3$$

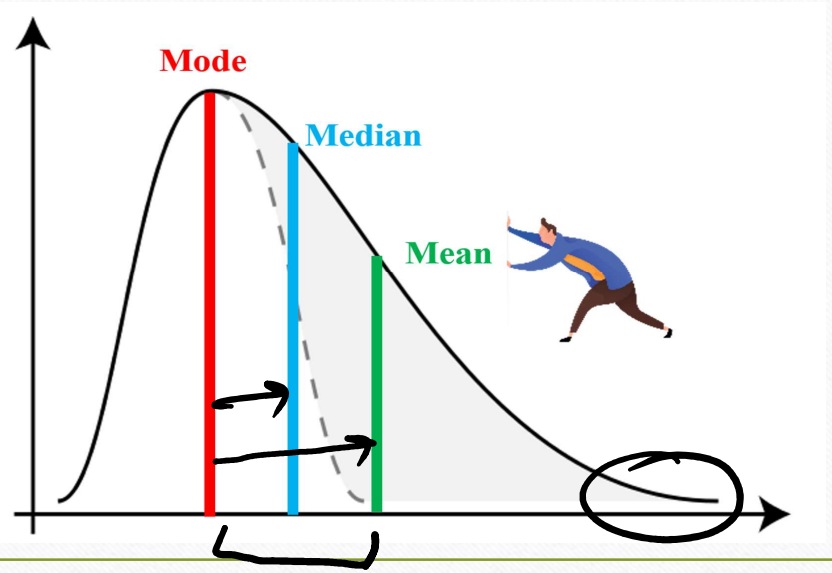


Skewness

Negative or Left Skewness



Positive or Right Skewness



'Karl Pearson'

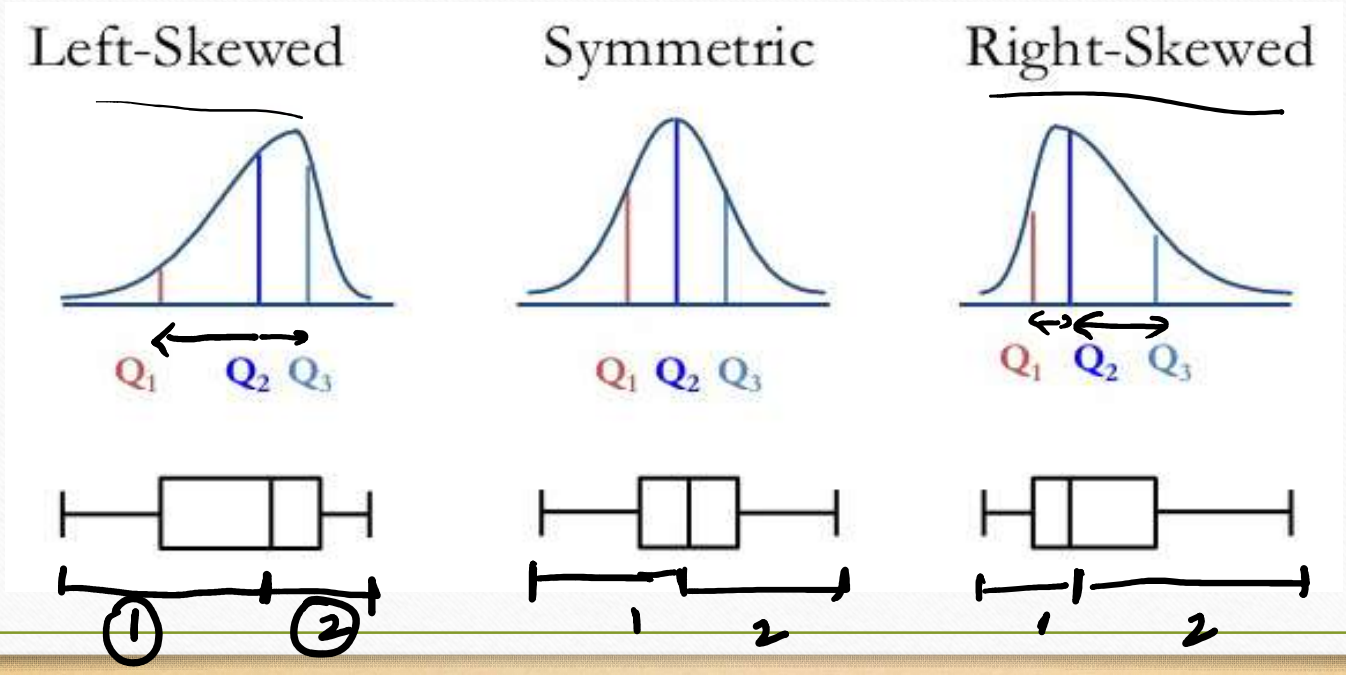
$$S_K = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

empirical formula: $\text{Mode} = 3\text{Median} - 2\text{Mean}$

$$S_K = \frac{\text{Mean} - (3\text{Median} - 2\text{Mean})}{\sigma}$$

$$S_K = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

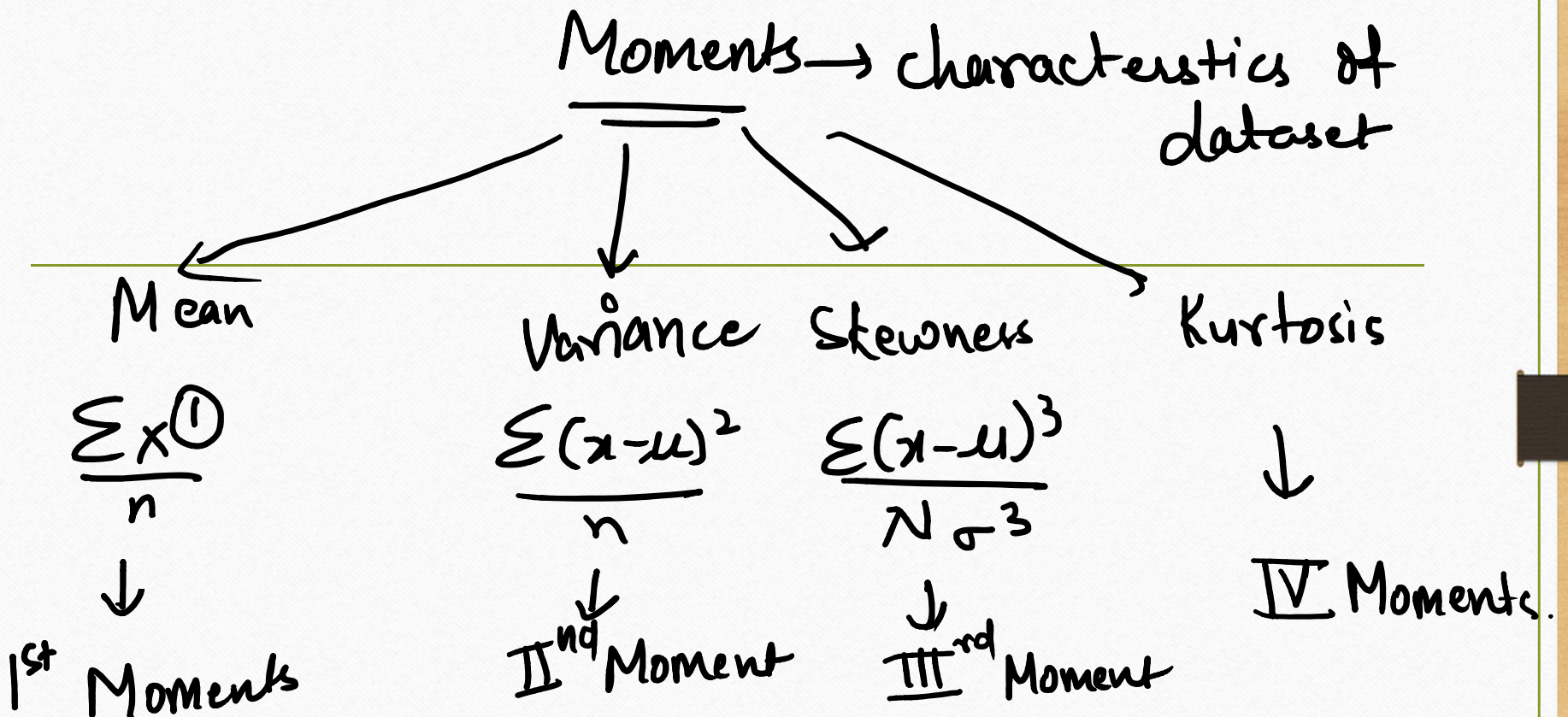
Skewness


Left-Skewed
Symmetric
Right-Skewed
Q1 Q2 Q3
Q1 Q2 Q3
Q1 Q2 Q3
1
2
1
2
1
2

① > ②

Code

```
import numpy as np
from scipy.stats import skew
x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution
print(skew(x))
```


Moments → characteristics of dataset
Mean
Variance
Skewness
Kurtosis
$\frac{\Sigma x^{(1)}}{n}$
$\frac{\Sigma (x-\mu)^2}{n}$
$\frac{\Sigma (x-\mu)^3}{N \sigma^3}$
1st Moments
IInd Moment
IIIrd Moment
IV Moments.

Stocks/Mutual funds

Low Risk

Mid risk

High risk

Kurtosis

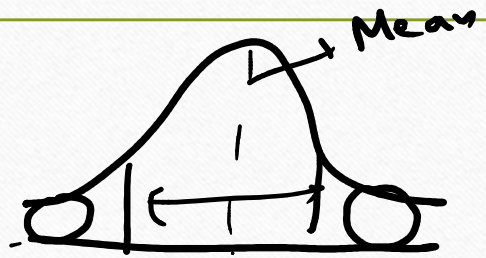
- Kurtosis is all about the tails of the distribution – not the peakedness or flatness. It measures the **tail-heaviness** of the distribution. If high, then data has lot of outliers.
- The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the **excess kurtosis** is presented: excess kurtosis is simply **kurtosis−3**. For example, the "kurtosis" reported by Excel or any statistical library is actually the excess kurtosis.

$K = 3$

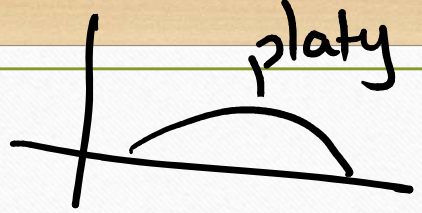
$$a_4 = \sum \frac{(X_i - \bar{X})^4}{ns^4}$$

(Kurtosis for ND)

excess kurtosis = 0 ⇒ kurtosis - 3



ek=0, k=3



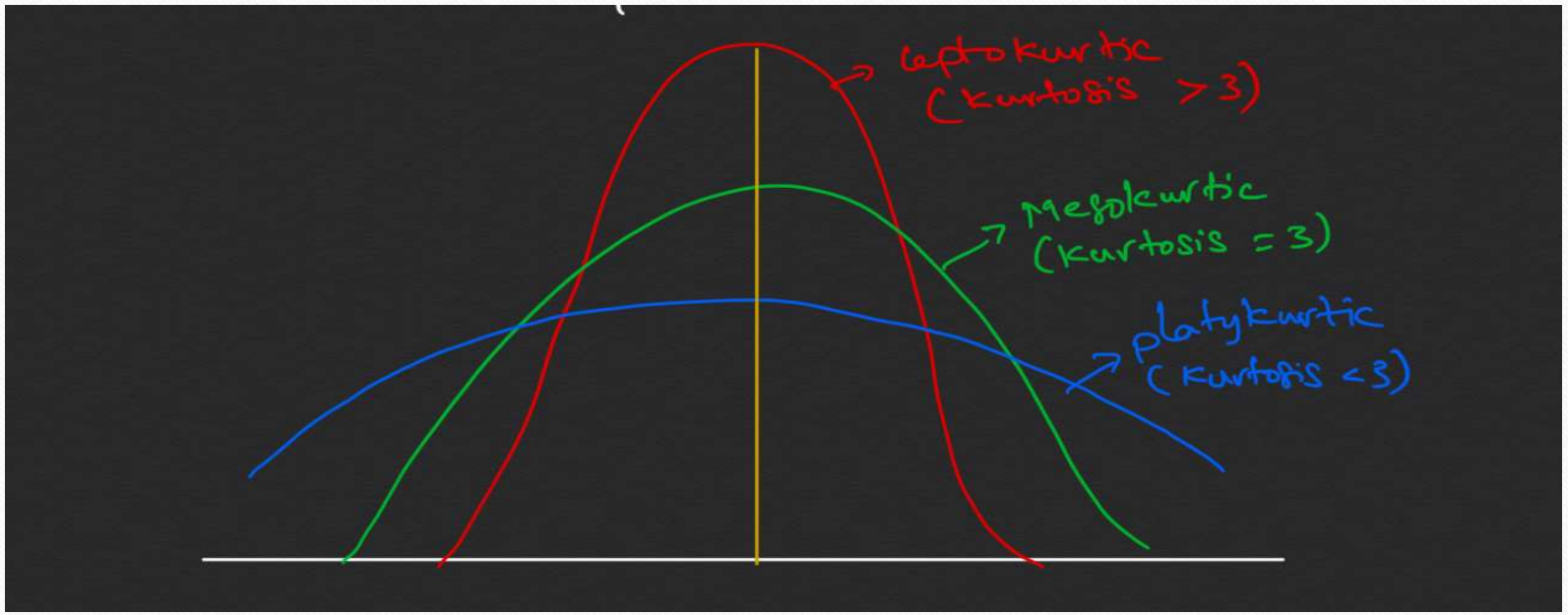
Kurtosis

ek<0, k<3

ek>0, k>3

- The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the **excess kurtosis** is presented: excess kurtosis is simply **kurtosis−3**. For example, the "kurtosis" reported by Excel or any statistical library is actually the excess kurtosis.
- 1. A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called **mesokurtic**.
- 2. A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- 3. A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

Kurtosis


leptokurtic
(kurtosis > 3)
mesokurtic
(kurtosis = 3)
platykurtic
(kurtosis < 3)

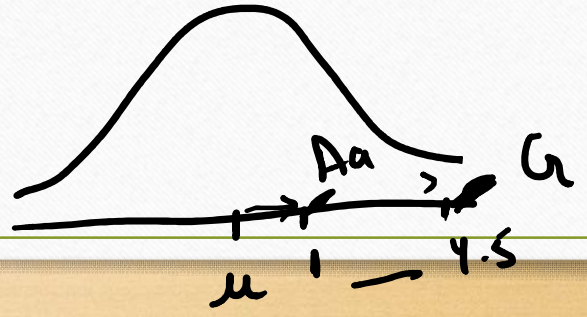
Code

```
import numpy as np

from scipy.stats import kurtosis

x = np.random.normal(0, 2, 10000) # create random values based on a normal distribution

print(kurtosis(x))
```

Z-Score Or Standard Score

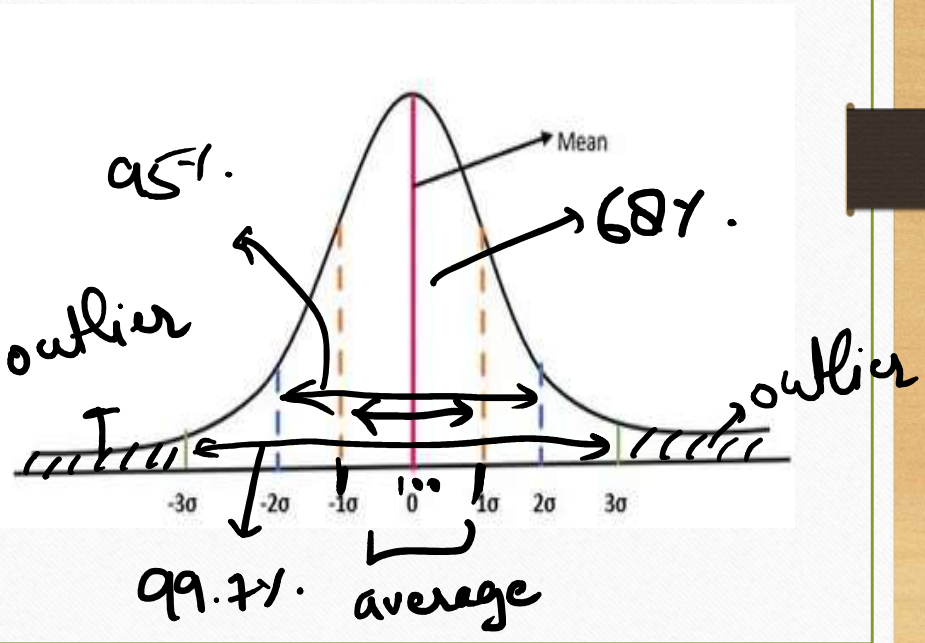
- Standard scores give you a way of comparing values across different sets of data where the mean and standard deviation differ. They're a way of comparing related data values in different circumstances.
- Standard score = number of standard deviations from the mean.
- The standard score makes such comparisons possible by transforming each set of data into a more generic distribution. We can find the standard score of each player at the practice session, and then transform and compare them.

$$\text{Z-Score} = \frac{x - \mu}{\sigma}$$



Z-Score Or Standard Score

- Standard scores work by transforming sets of data into a new, theoretical distribution with a mean of 0 and a standard deviation of 1.
- Standard scores can take any value, and they indicate position relative to the mean. Positive z-scores mean that your value is above the mean, and negative z-scores mean that your value is below it. If your z-score is 0, then your value is the mean itself. The size of the number shows how far away the value is from the mean.



$R = 5km$ Avg Rod $= 1.5km$

$\sigma = 1$

$$Z_R = \frac{x - \mu}{\sigma} = \frac{5 - 1.5}{1} = 3.5$$

Milk $= 2l$, $\sigma = 1$

Avg Milk $= 0.5l$

$$Z_M = \frac{2 - 0.5}{1} = 1.5$$