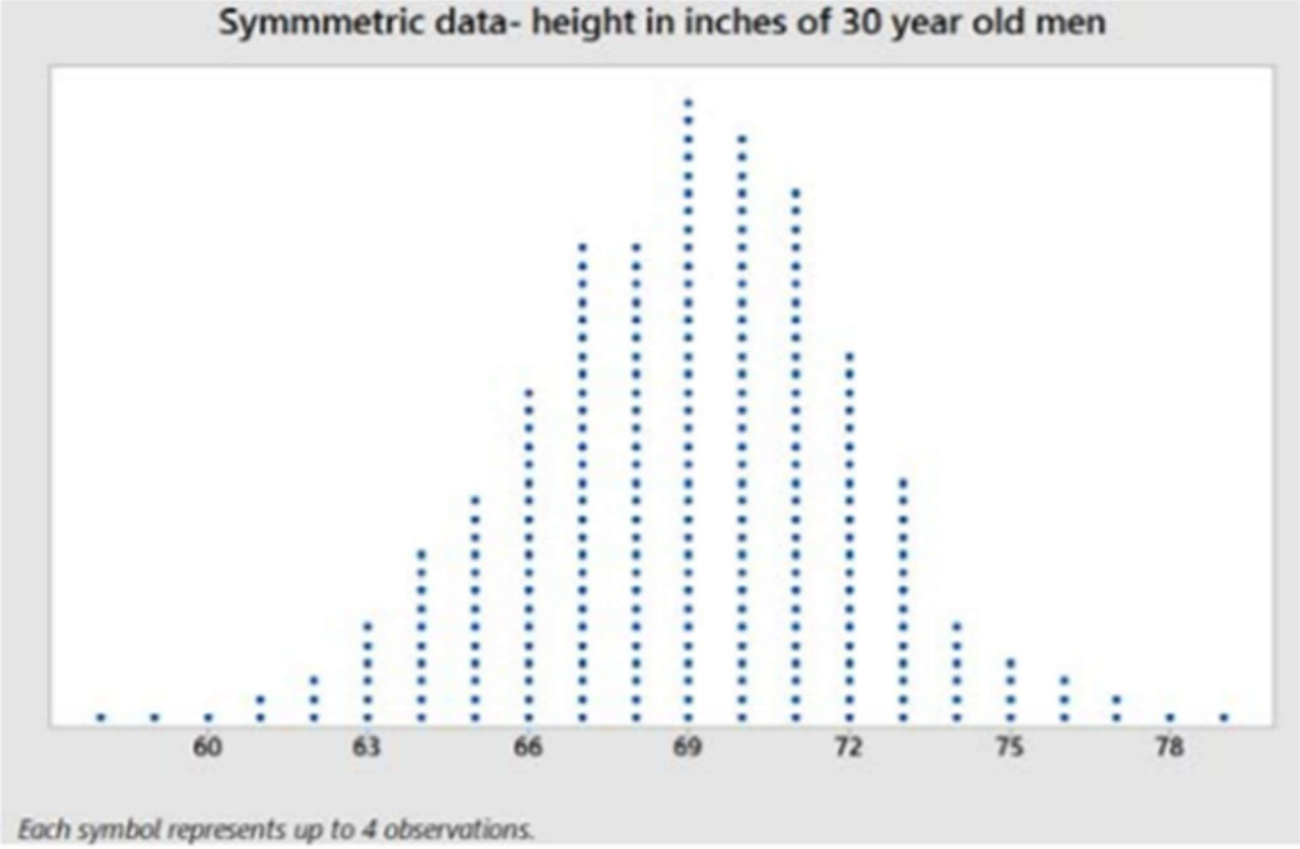
Mithilesh Singh

Using the mean and median to determine skewness

Skewness is a measure of how asymmetric the data vales are. Data can be positively skewed (stretched to the right), negatively skewed (stretched to the left) or symmetric (no skewness). Let's now explore what effect skewness has on measures of center with several examples.

Here is a dot plot and summary statistics of the heights in inches of 1000 men, aged 30 years



Sample mean = 68.98 inches

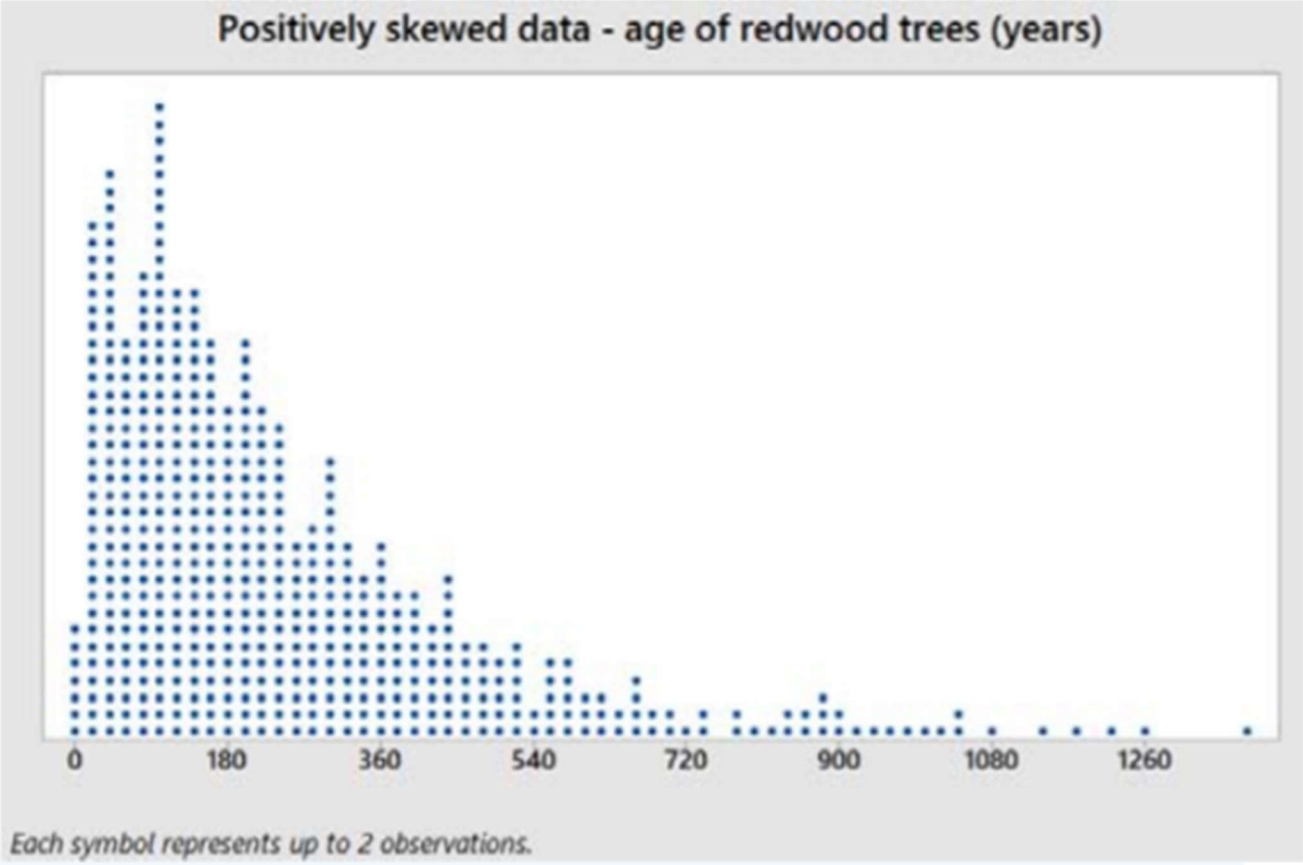
Sample median = 69 inches

Sample mode = 69 inches

The data values are evenly spread on the right and left of the peak. When data are symmetric, the mean, median and mode are about the same.

Example: Positively skewed data – Redwood trees

Here is a dot plot and summary statistics of the age of 1000 redwood trees sampled in California parks.



Sample mean = 237.48 years

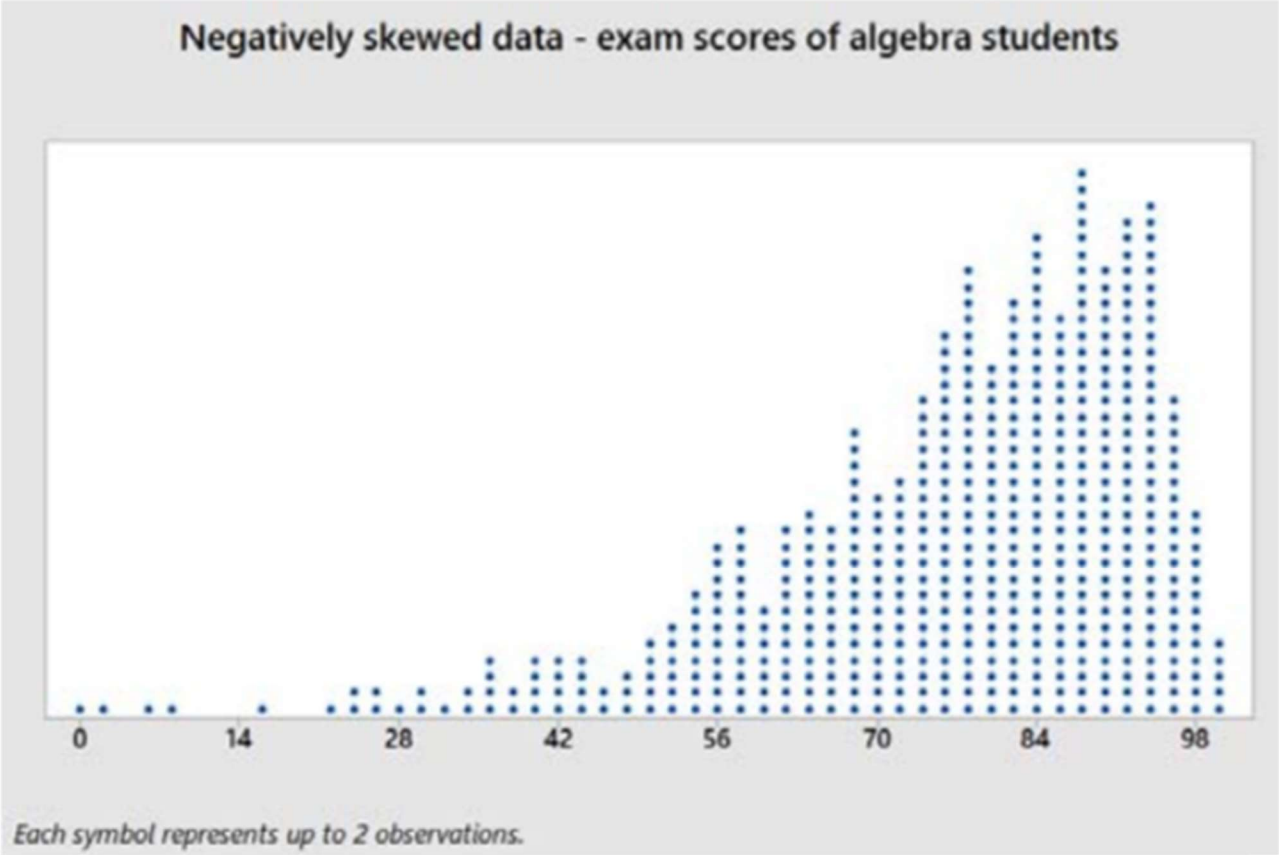
Sample median = 180 years

Sample mode = 100 years

The data values are stretched to the right of center, causing the mean to be greater than the median. Also, the median will usually be greater than the mode for positively skewed data.

Example: Negatively skewed data – Exam grades

Here is a dot plot and summary statistics of the percentage grade of 1000 midterm exams given by a math instructor to algebra students.

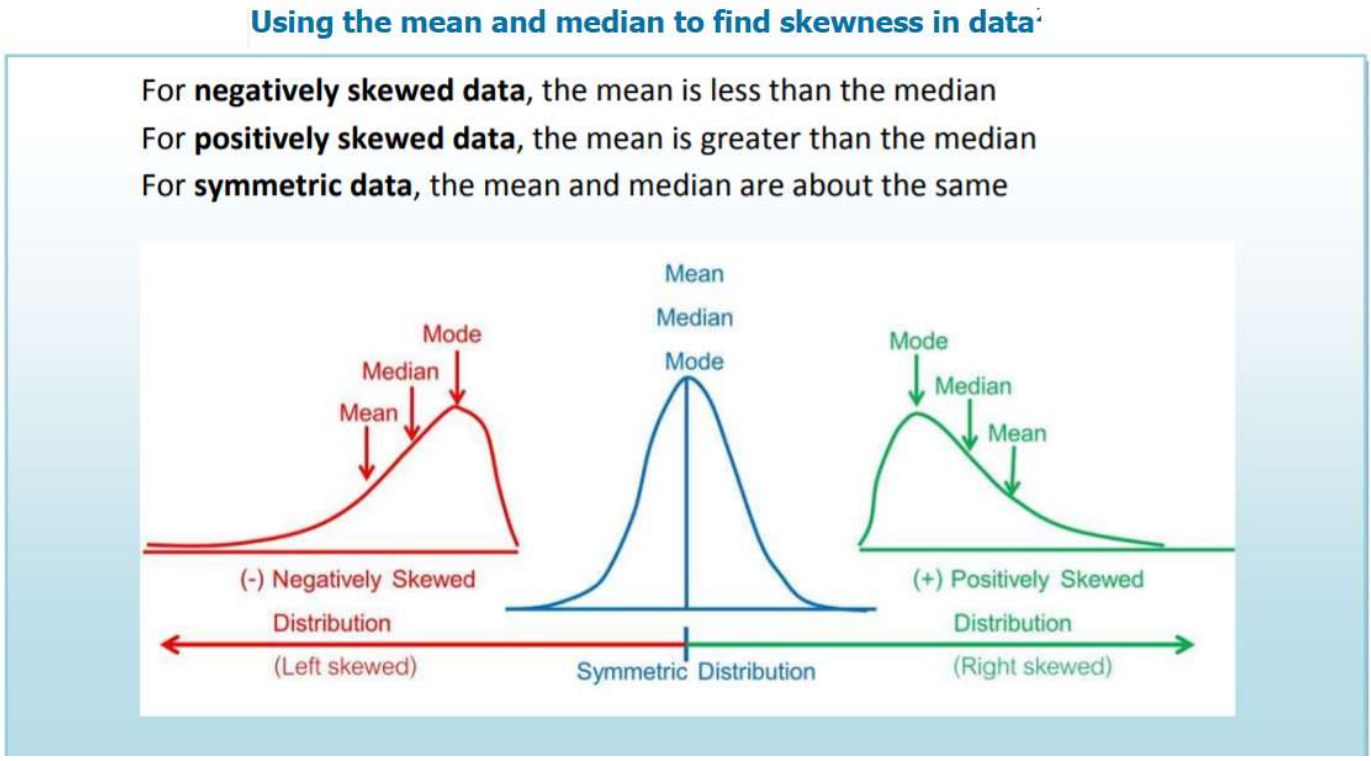


Sample mean = 76.21

Sample median = 80

Sample mode = 91

The data values are stretched to the left of center, causing the mean to be less than the median. Also, the median will usually be less than the mode for negatively skewed data.



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. Skewness is calculated as:

```
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))
#using on dataframe
df.skew()
```

Degree of skewness

In real world scenario we will not get a perfect symmetrical data

Rule of thumb

If the skewness is between –1 to 1 data is fairly symmetrical

5-point summary

Quantile

(MINIMUM)

Q1- first quantile

Q2-Median (2^{nd} quantile)

Q3- third quantile

(MAXIMUM)

After arranging the dataset in ascending order, when we find the position of the median, the position of median is dividing the dataset in to two group.

Q2= Median of the complete dataset. It divides the dataset in two equal group.

Q1= Median position of the first half is first quantile.

Q3= Median position of the second half is Third quantile.

Inter Quantile Range

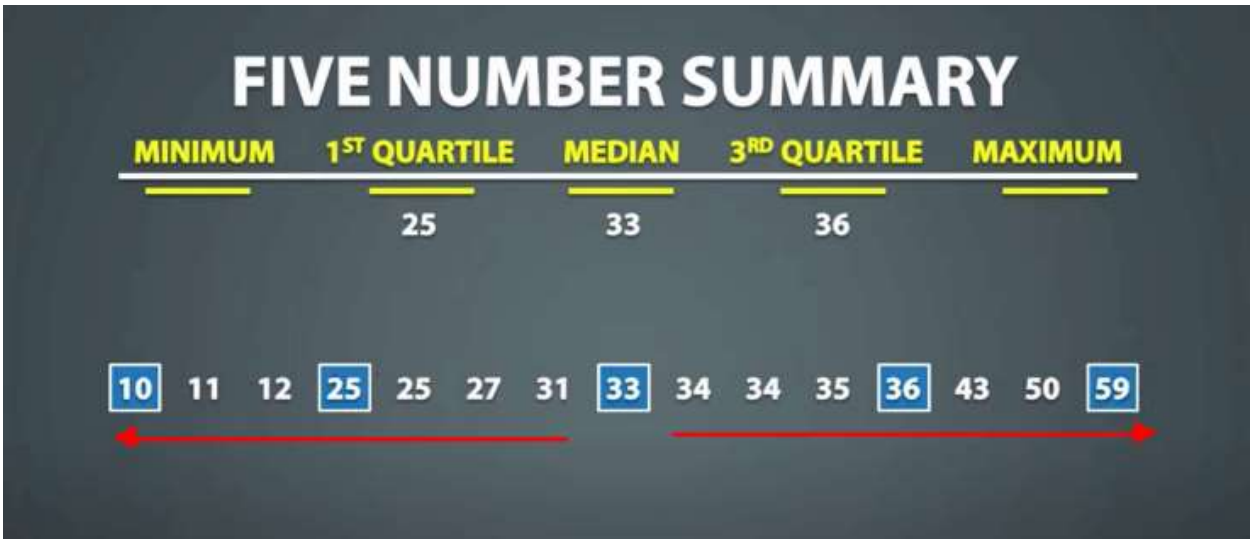
IQR=Q3-Q1.

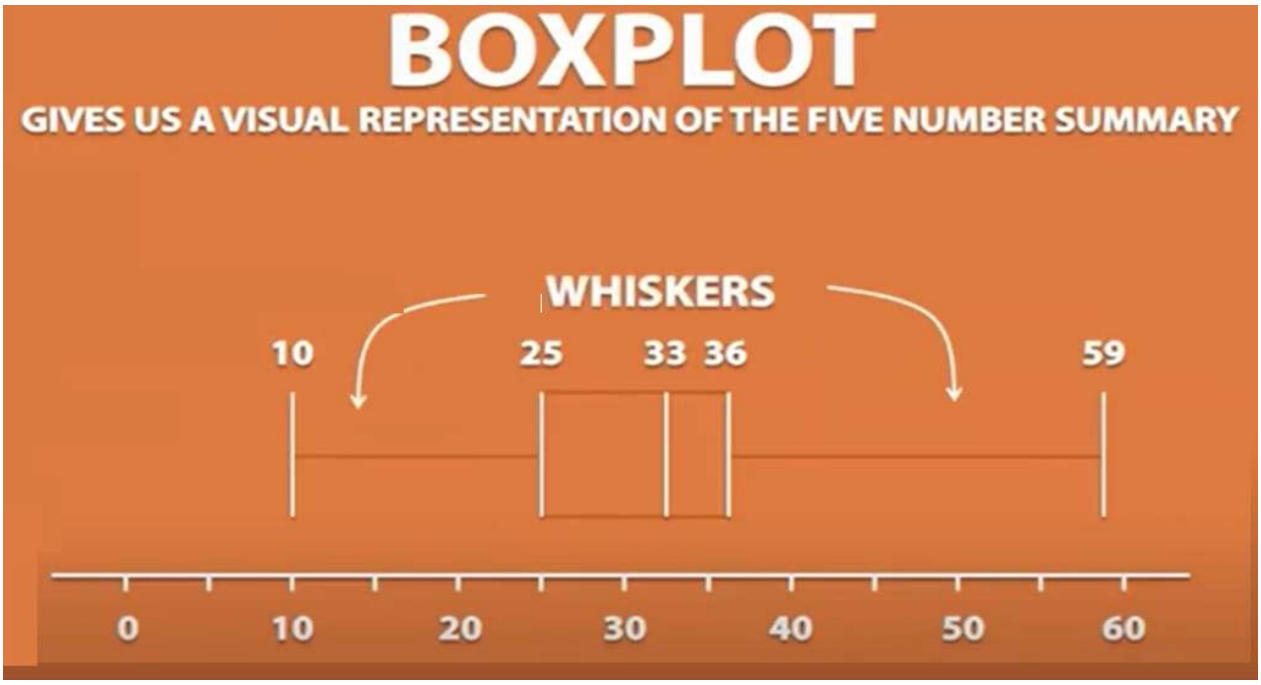
Find Q1, Q2, Q3, IQR?

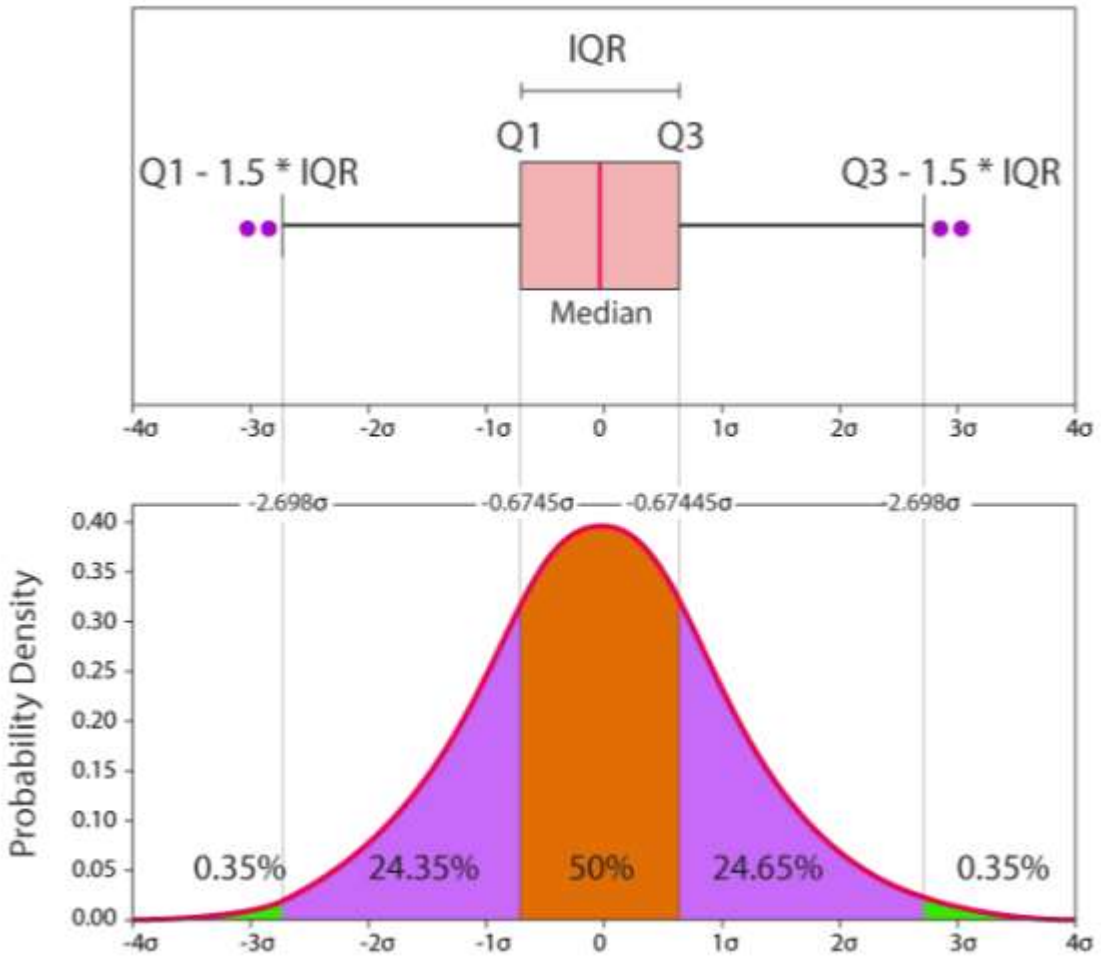
12 25 25 27 10 11 31 33 36 43 50 59 34 34 35

Data point arranged in ascending order,

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59



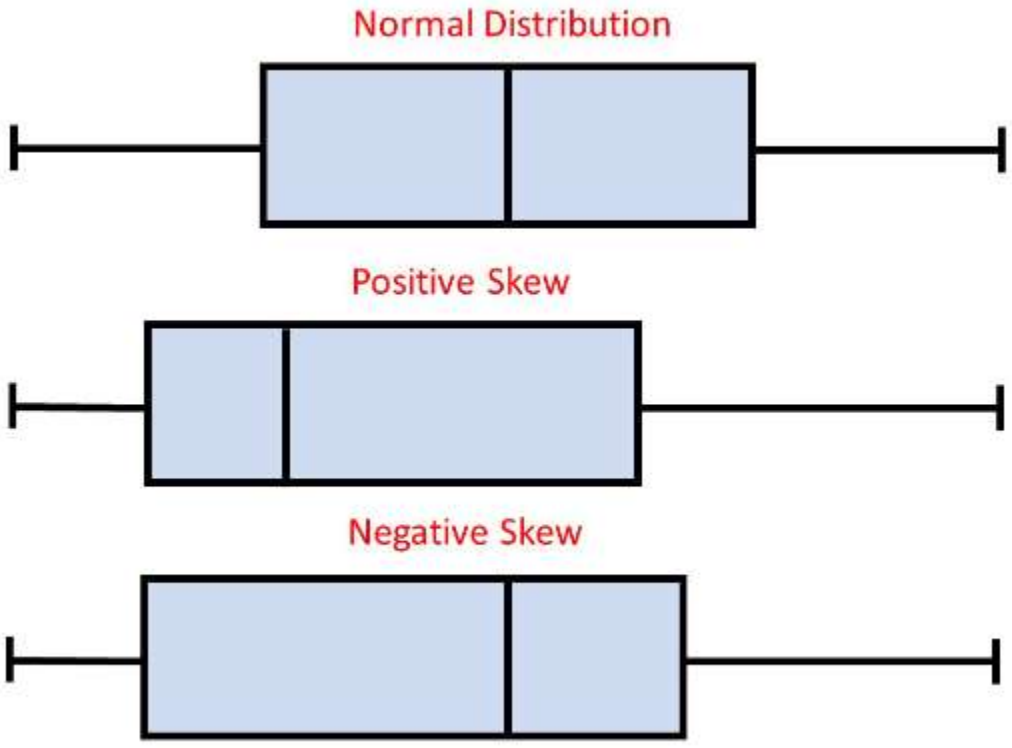


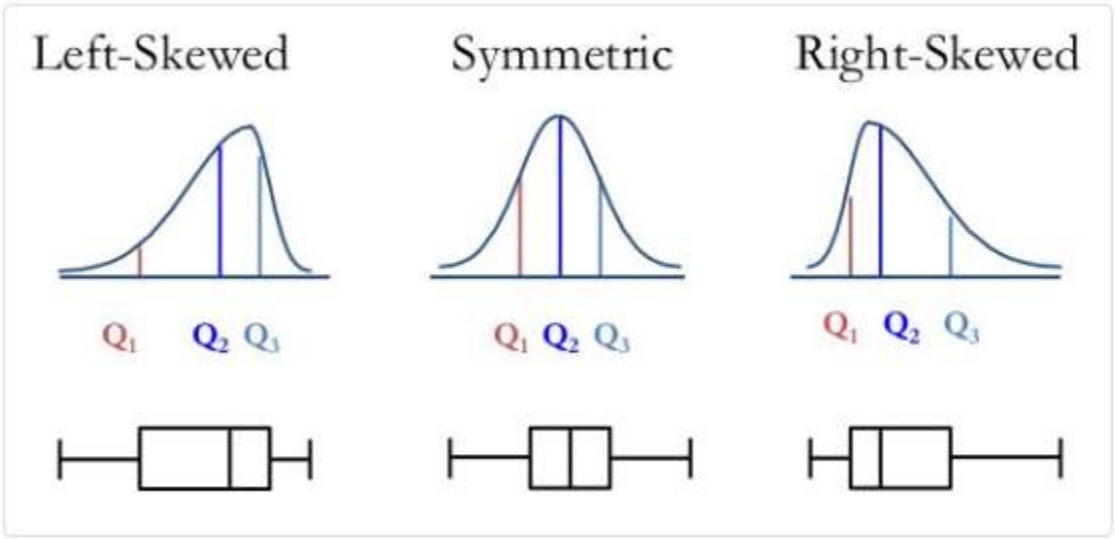


Boxplot on a normal distribution

Also called Gaussian Distribution

Similar to bell shape curve….

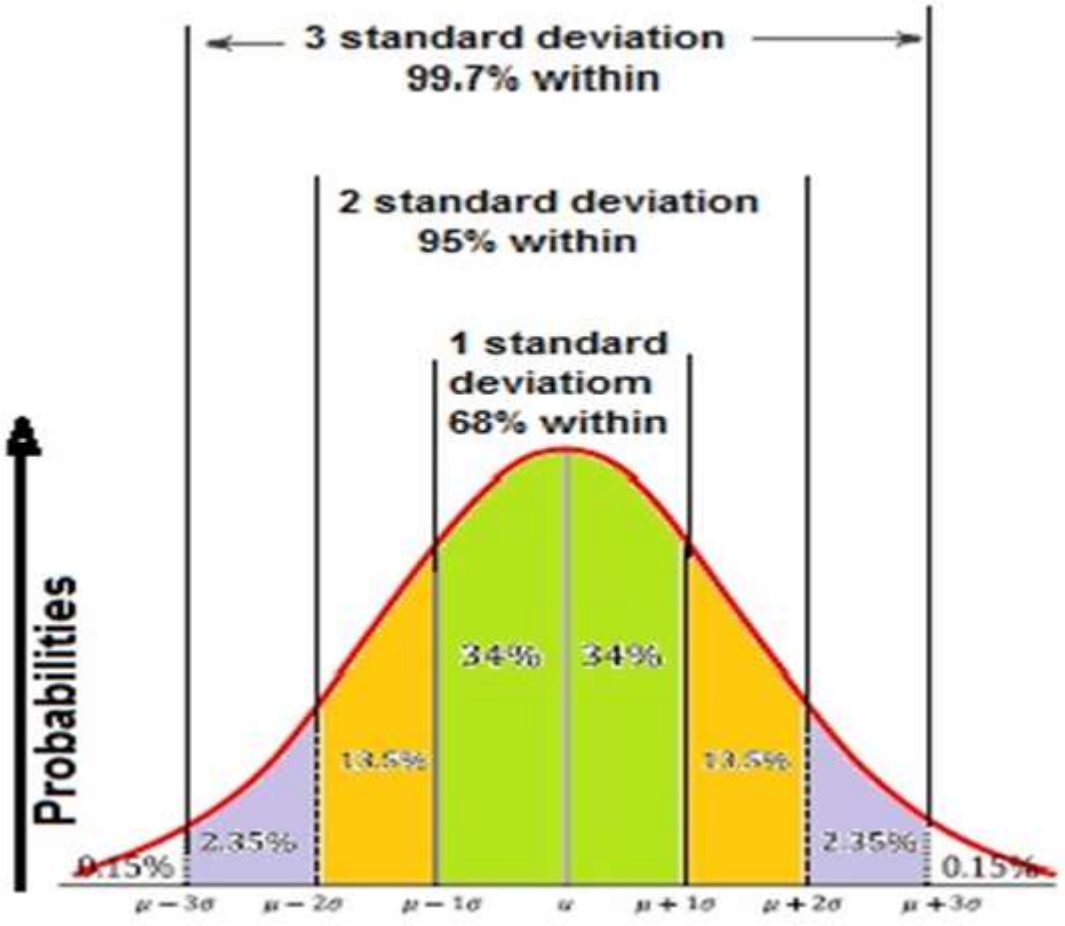




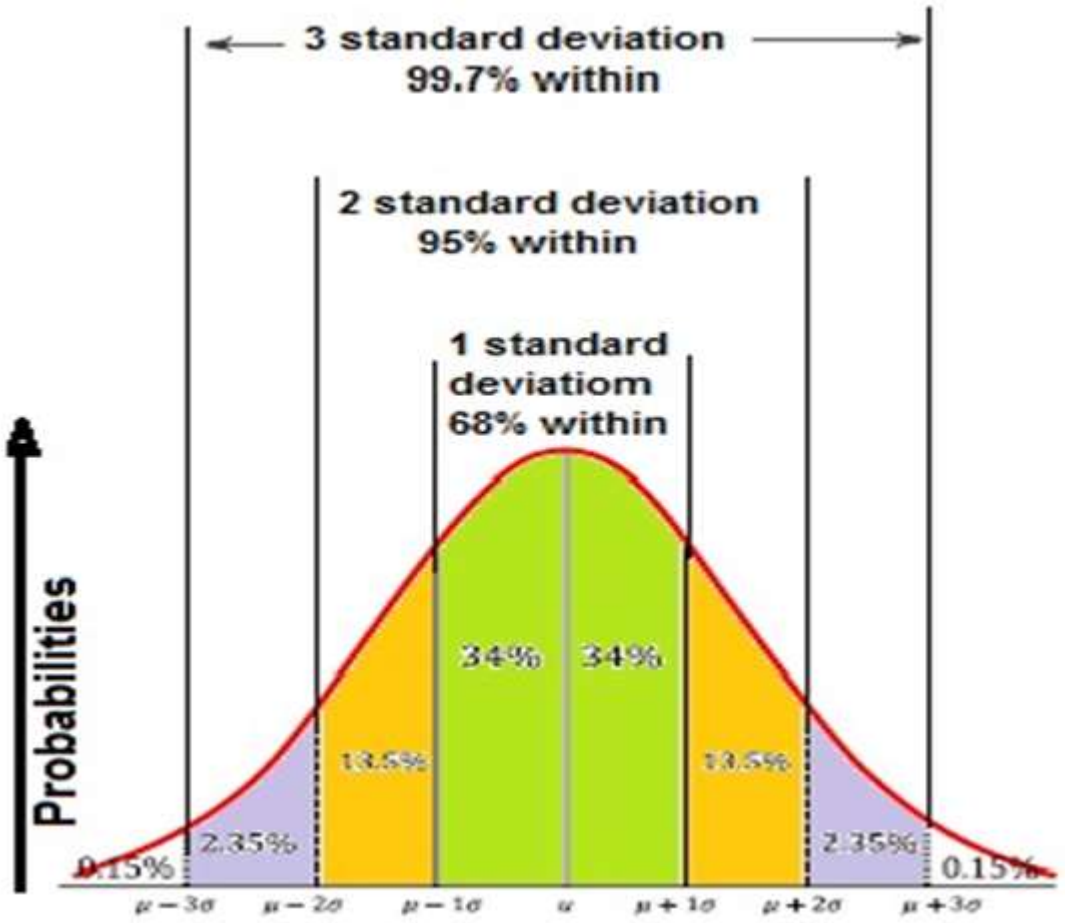
Empirical Rule.

BELL shape and symmetric.

A useful continuous distribution is the standard normal distribution with mean equal to 0 and standard deviation equal to 1



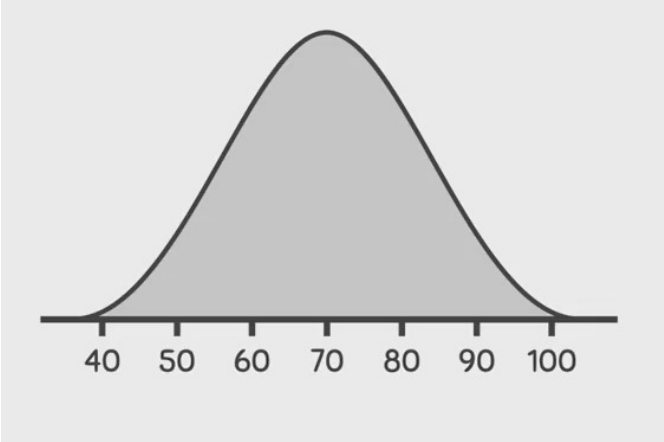
Normal distribution is a density curve total area equals to 100%.



There are certain observations which could be inferred from this Rule 68,95,99. Called Empirical Rule.

About 68.26% data lies within 1 SD (<σ) of the mean (μ), About 95.44% data lies within 2 s d (<σ) of the mean (μ), About 99.72% data lies within 3 s d (<σ) of the mean (μ) & the rest 0.28% of the data lies outside 3 SD (>3σ) of the mean (μ), And this part of the data is considered as outliers.

(1) The Normal distribution has a standard deviation of 10 and mean 70. Approximately what area is contained between 70 and 90?



Hint:- use Empirical Rule 68,95,99.

(2) Q. In 2019 base salary of NYC employees between $ 1000 and $ 150000.

Mean =$ 73,555.88

SD=$ 27,505.98

What proportion of population makes at most $ 128,567.80?

(3) **what proportion of population makes more than $ 101,061.90?**

Mean =$ 73,555.88

SD=$ 27,505.98

(4) In a Garden the heights of plants are normally distributed, the mean of the plants are 22.2 inches and the SD of 4.5 inches.

Estimate the percentage of plants that are less than 13.2 inches tall.

Hint: - Use Empirical rule.

Q.5 . The heights of adult women are normally distributed with a mean of 62.5 inches and standard deviation of 2.5 inches. Use the Empirical Rule to determine between what two heights 68% of adult women will fail.

A. (52.5,72.5)
B. (55,70)
C. (57.5,67.5)
D. (60,65)

Q6 . The heights of adult women are normally distributed with a mean of 62.5 inches and standard deviation of 2.5 inches. Use the Empirical Rule to determine between what two heights 99.7% of adult women will fail.

A. (52.5,72.5)
B. (55,70)
C. (57.5,67.5)
D. (60, 65)

Q.7 Variable A is normally distributed with μ = 12.00 and σ = 3.11. What is the probability that a randomly selected case will have a score of less than 15?

(A)0.72
(B)0.29
(C)0.87
(D)0.12

Q.8 The shelf life of a particular dairy product is normally distributed with a mean of 12 days and a standard deviation of 3 days.
About what percent of the products last between 12 and 15 days?
(A)68%
(B)34%
(C)16%
(D)2.5%

Q.9 A machine produces electrical components.
99.7% of the components have lengths between 1.176 cm and 1.224 cm. Assuming this data is normally distributed, what are the mean and standard deviation?

(A)Mean = 1.210 cm
S.D. = 0.008 cm
(B)Mean = 1.200 cm
S.D. = 0.004 cm
(C)Mean = 1.190 cm
S.D. = 0.008 cm
(D)Mean = 1.200 cm
S.D. = 0.008 cm