Basic Statistics for Data Science-2

Arif Istiake Sunny

sunny1509006@gmail.com

Agenda

- Standard Daviation
- Normal Distribution
- Deciles, Percentiles and Quartiles
- Inter-quartile Range
- Five Number Summary
- Box Plot
- Coefficient of Correlation
- Standard Score (Z-score, T-score)
- Population Data Vs Sample Data

Standard Deviation (আদর্শ বিচ্যুতি)

				(Mean)			
Result	**38**	**44**	**45**	**50**	**55**	**44**	**74**

From which number we can say it is Good or Bad result ?

Mean plus or minus standard deviation is optimum result but beyond this is either good or bad

Suppose SD = 2

Then optimum number is 50+2 = 52 or 50 -2 = 48

Standard Deviation (আদর্শ বিচ্যুতি)

How to calculate the Standard Deviation

2, 8, 2, 8 Mean = 5

$$\sigma^2 \text{ Variance} = \frac{(2-5)^2 + (8-5)^2 + (2-5)^2 + (8-5)^2}{4} = 9$$

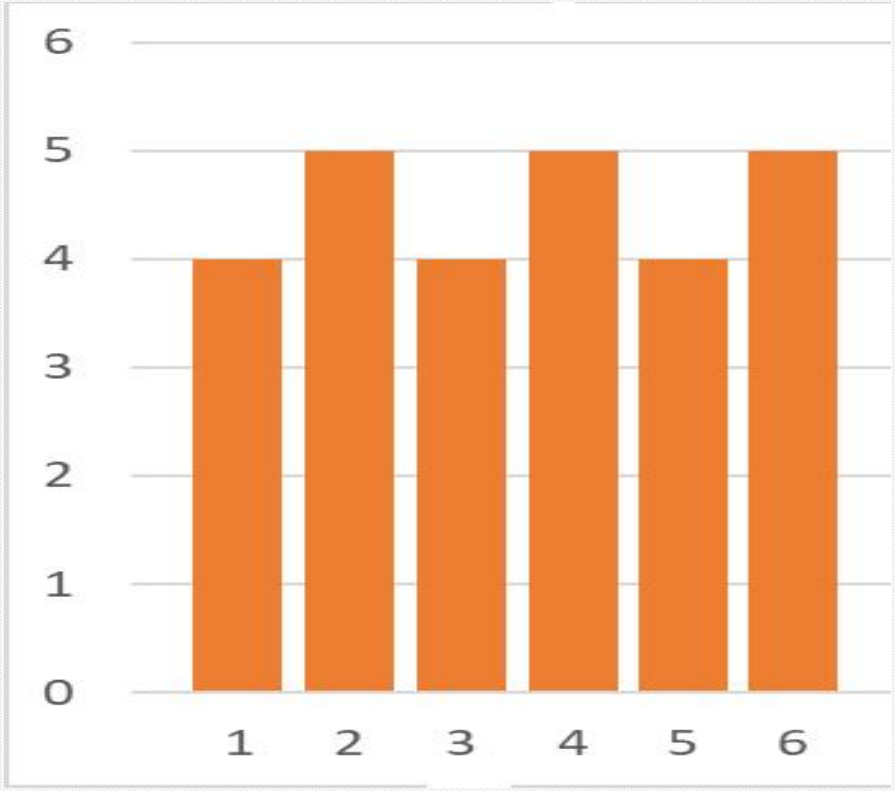
$$\sigma = \sqrt{\sigma^2} \qquad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{9} = 3$$

Normal Distribution
Gaussian Distribution
(পরিমিত বিন্যাস)

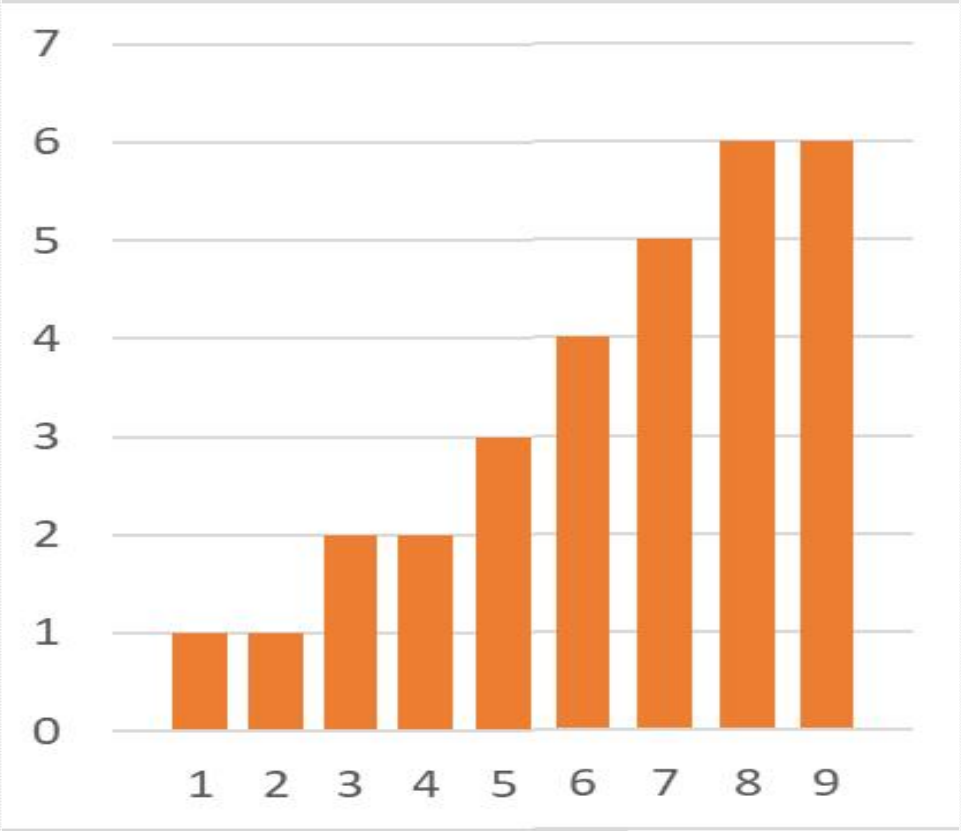
Uniform Distribution

[1,1,1,1, 2,2,2,2,2, 3,3,3,3, 4,4,4,4,4, 5,5,5,5, 6,6,6,6,6]



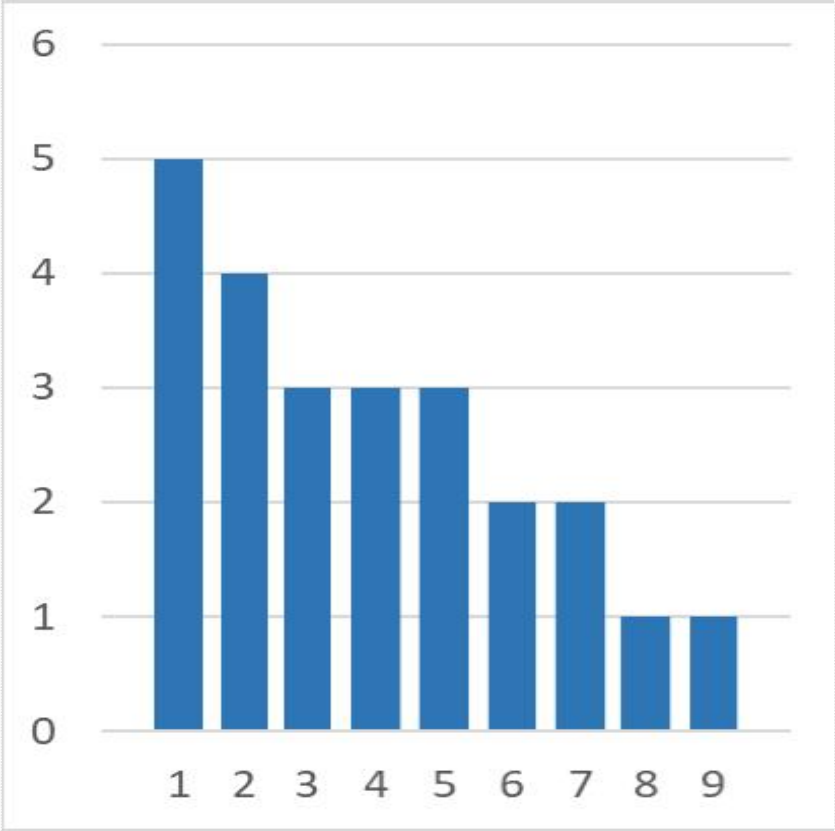
Left Skewed Distribution

[1, 2, 3,3, 4,4, 5,5,5, 6,6,6,6, 7,7,7,7,7, 8,8,8,8,8,8, 9,9,9,9,9,9]



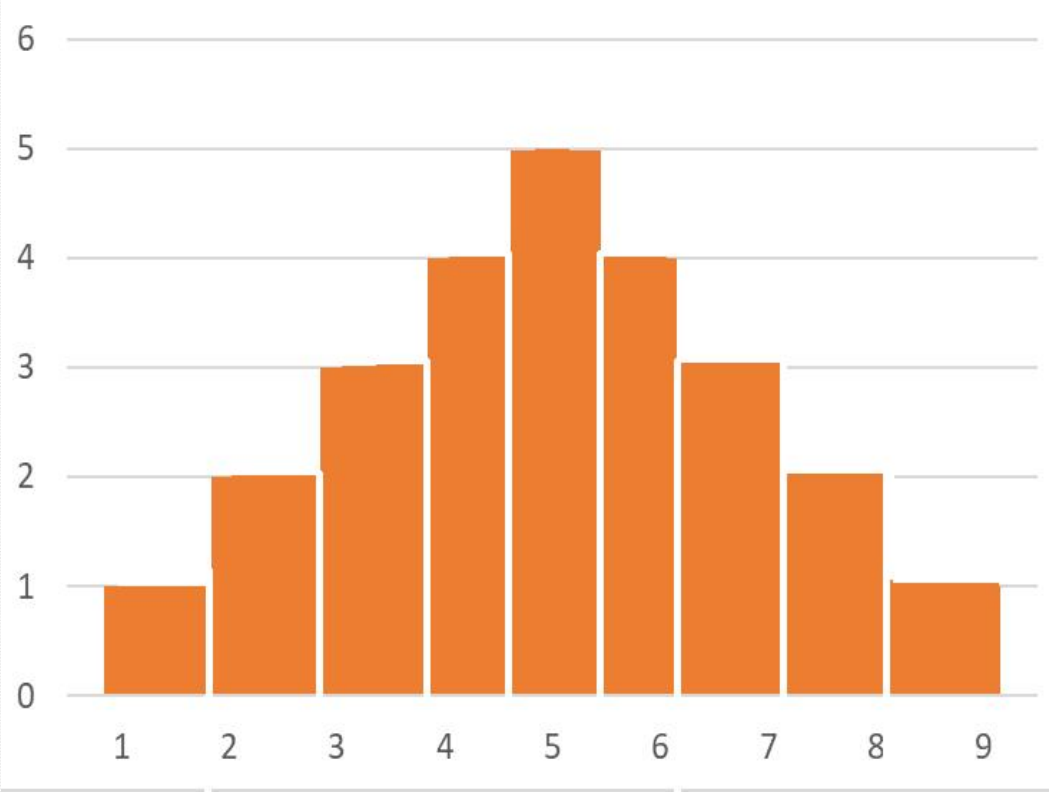
Right Skewed Distribution

[1,1,1,1,1, 2,2,2,2, 3,3,3, 4,4,4, 5,5,5, 6,6, 7,7, 8, 9]

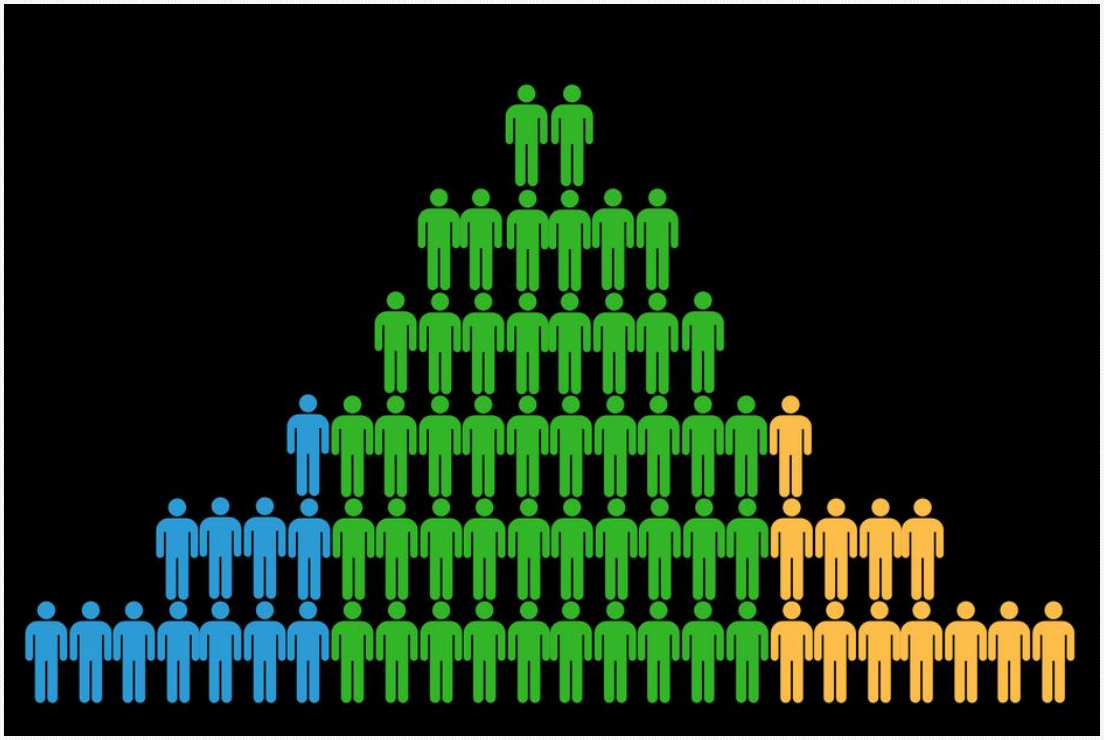


Normal Distribution/Gaussian distribution

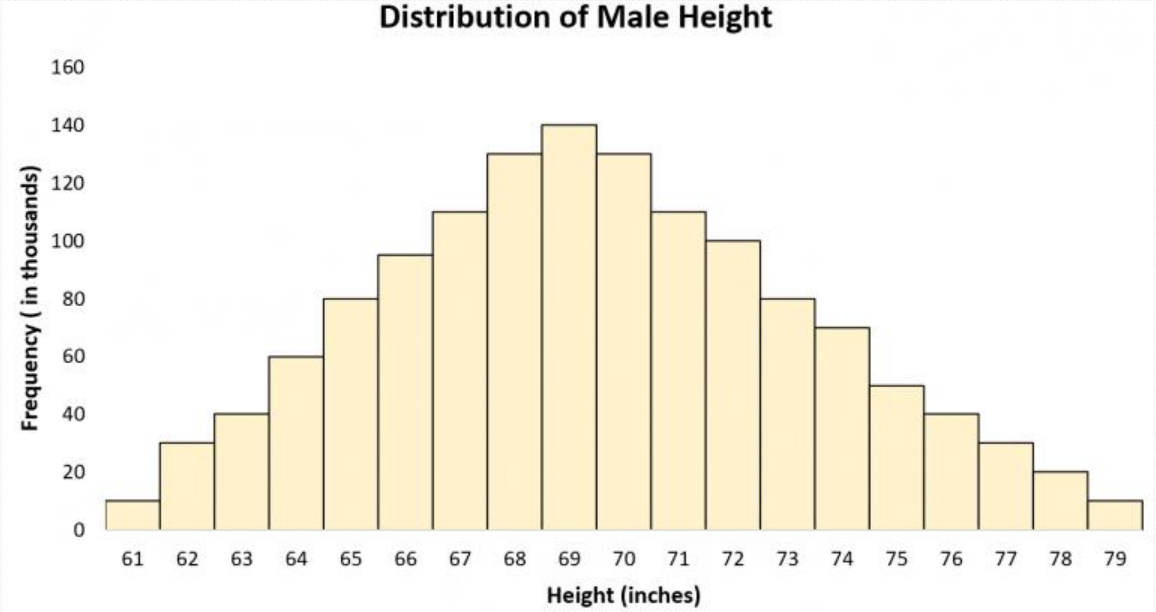
[1, 2,2, 3,3,3, 4,4,4,4, 5,5,5,5,5, 6,6,6,6, 7,7,7, 8,8, 9]



Normal Distribution/Gaussian distribution

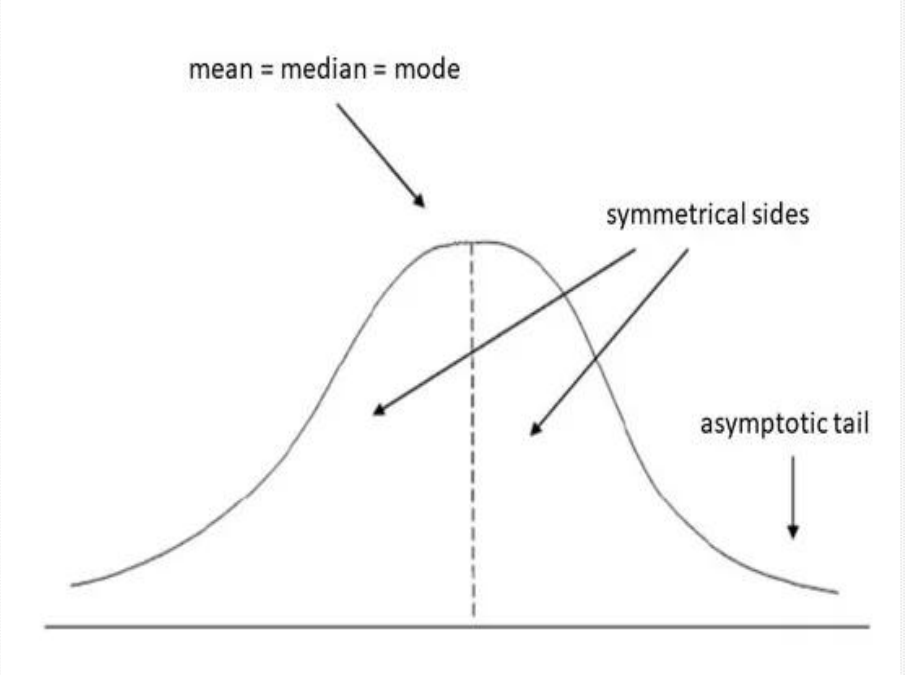


Normal Distribution



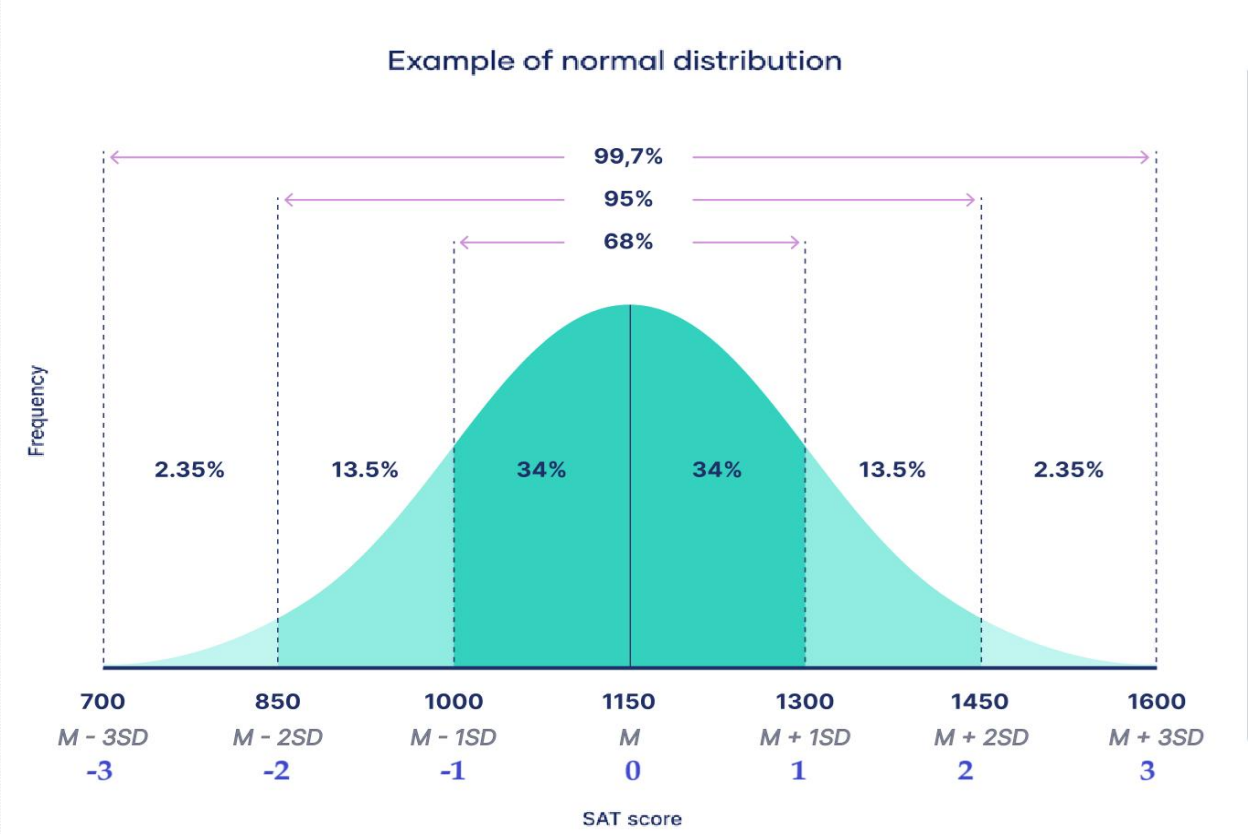
Normal Distribution

- Bell Curve
- Mean = Median = Mode
- Symmetrical
- Kurtosis = 0.263
- Asymptotic
- SD as a unit of Measurement
- -3SD to +3SD

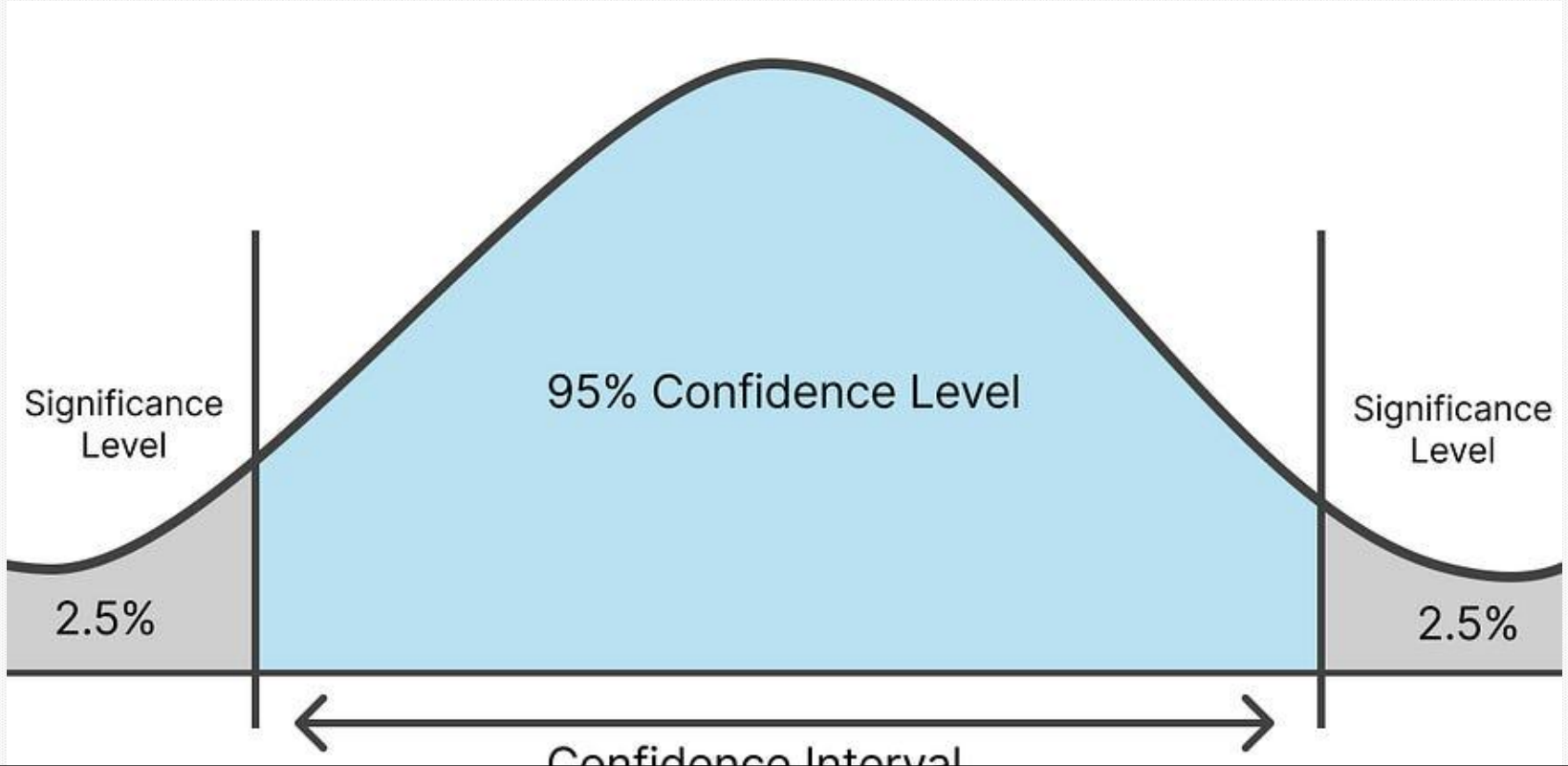


Normal Distribution/Gaussian distribution

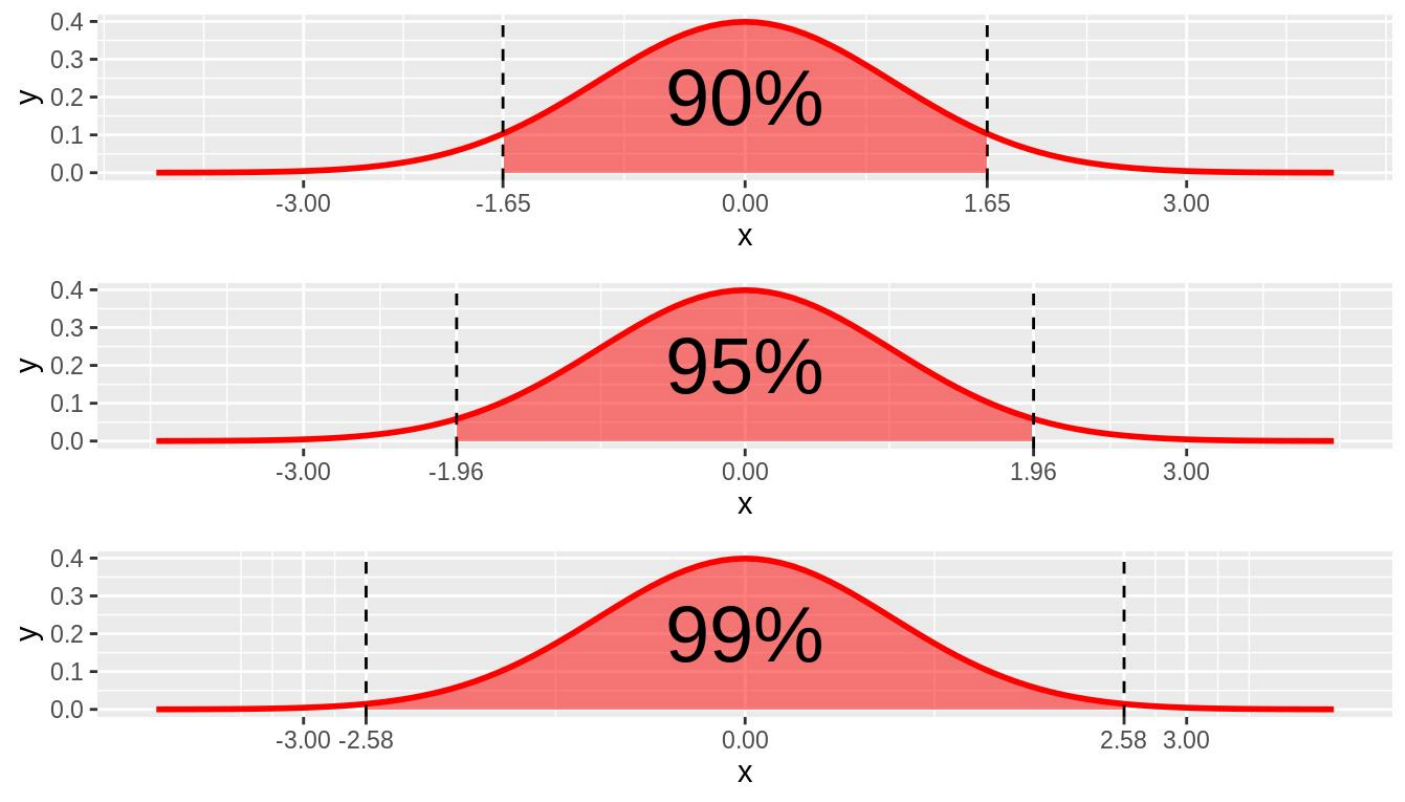
- Bell Curve
- Mean = Median = Mode
- Symmetrical
- Kurtosis = 0.263
- Asymptotic
- Mean as a starting point
- SD as a unit of Measurement
- -3SD to +3SD



Confidence Level/Significant Level



Confidence Level/Significant Level



Normal Distribution/Gaussian distribution

নরমাল বিন্যাস-এর ব্যবহার

- অভিজ্ঞতায় দেখা গিয়েছে মানুষের উচ্চতা, ভর (ওজন) ইত্যাদি নরমাল বিন্যাস অনুসরণ করে।
- পুরুষদের উচ্চতার বিন্যাস নরমাল হলে–
  - বাংলাদেশের একজন পুরুষকে যদি দৈব চয়নের মাধ্যমে নির্বাচন করা হয় তাহলে তার উচ্চতা ৫ ফুট থেকে ৫.৫ ফুটের মধ্যে হবে তার সম্ভাবনা কত?
  - কিংবা সেই লোকটির ভর (ওজন) ৬০ থেকে ৭০ কিলোগ্রামের মধ্যে হবে তার সম্ভাবনা কত?
- জনমত জরিপে দেখা গেল ৫৫% উত্তরদাতা নগরসেবা বাড়ানো হবে এই শর্তে কর বাড়ানোর পক্ষে মত দিয়েছে।
  - জরিপ থেকে প্রাপ্ত এই শতকরা মানটি প্রকৃতপক্ষে (অর্থাৎ সেই জনগোষ্ঠীর সকলের মতামত যদি নেয়া যেতো তাহলে) ৪৫-৬৫% এর মধ্যে --তার সম্ভাবনা কত?
- নির্বাচনের আগে প্রধান দুই রাজনৈতিক দল নানা সংস্থাকে দিয়ে জরিপ পরিচালনা করে থাকে। একটি দল জরিপ করে জেনেছে তাদের পক্ষে ৪৫% জনসমর্থন আছে।
  - জনগোষ্ঠীতে এই সংখ্যা ৪২-৪৮% এর মধ্যে হবে তার সম্ভাবনা কত?
  - জনগোষ্ঠীতে এই সংখ্যা ৪০% এর কম তার সম্ভাবনা কত?

Deciles and Percentiles

Deciles: **If data is ordered and divided into 10 parts, then cut points are called Deciles**

Percentiles: **If data is ordered and divided into 100 parts, then cut points are called Percentiles.** 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Methods of Variability Measurement

Quartiles: **If data is ordered and divided into 4 parts, then cut points are called Quartile**

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25^{th} and 50^{th} percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

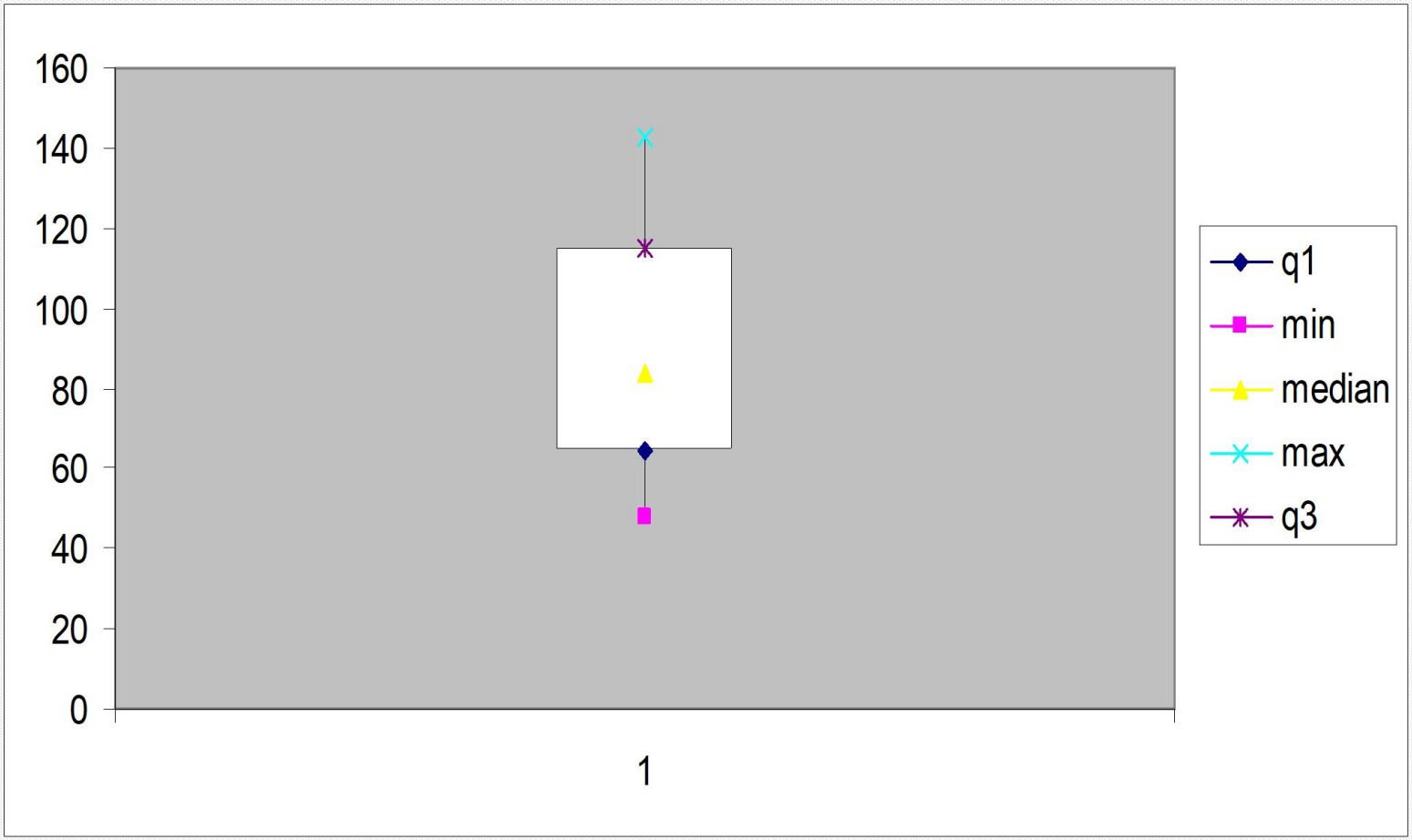
Inter-quartile Range: **Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 40=21. The middle half of the ordered data lie between 40 and 61.**

Five Number Summary

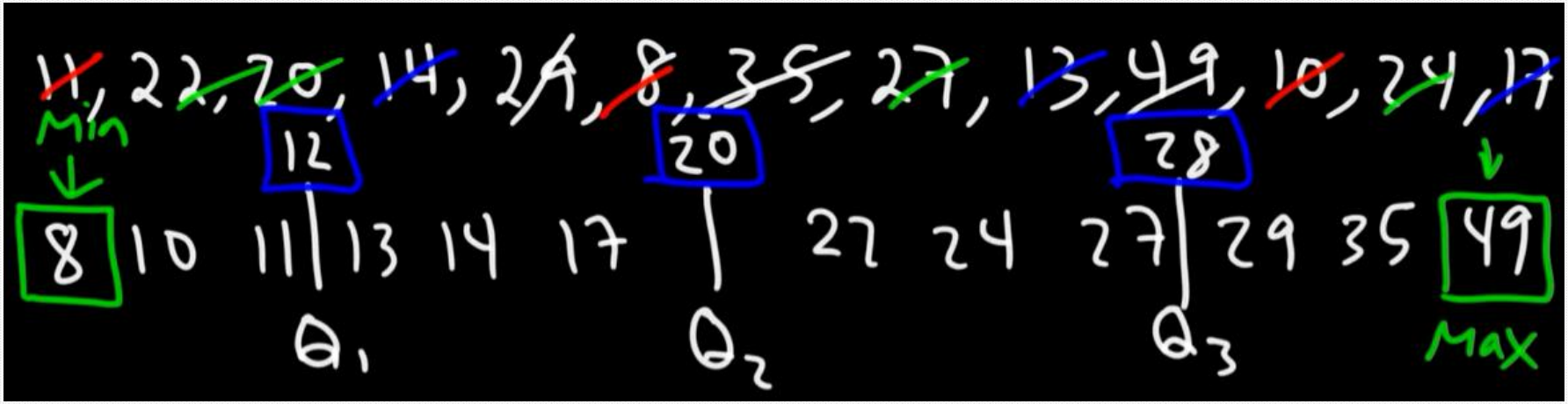
Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), The median(Q2), the third quartile(Q3), and the largest (Maximum) observation written in order from smallest to largest.

Box Plot: **A box plot is a graph of the five number summary.** The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

Boxplot



Boxplot



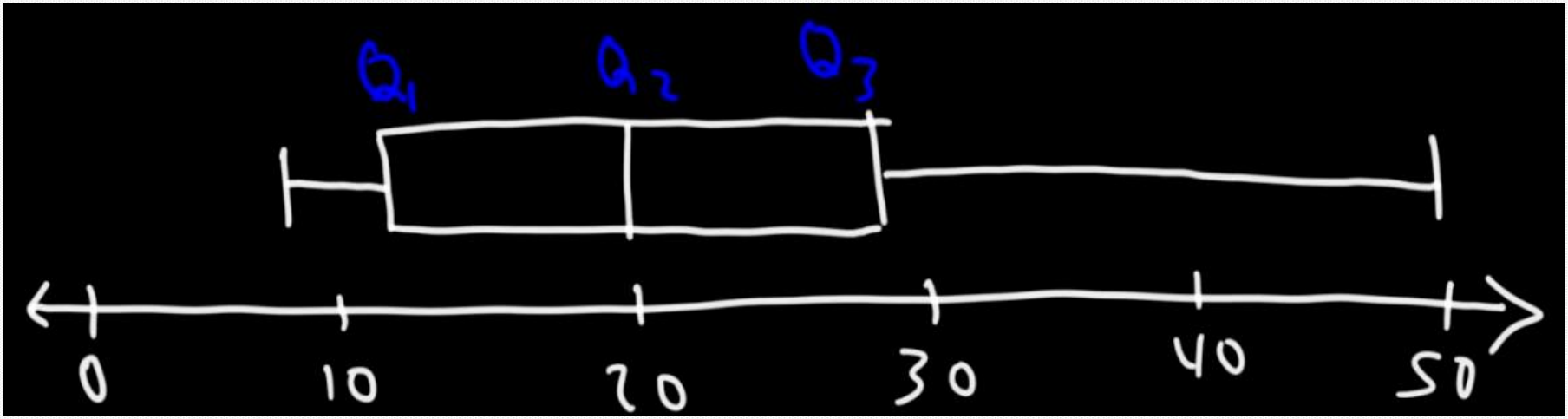
outlier

[Q1 - 1.5*IQR, Q3 + 1.5*IQR]
= [12 - 1.5*16, 28 + 1.5*16]
= [12-24, 28+24] = [-12, 52]

IQR = Q3-Q1 = 28-12=16

Above or under this range is outlier

Boxplot



18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 4

22 33 43

15 18 20 25 29 32 34 38 41 46 54 76

Q1 Q2 Q3

Outlier

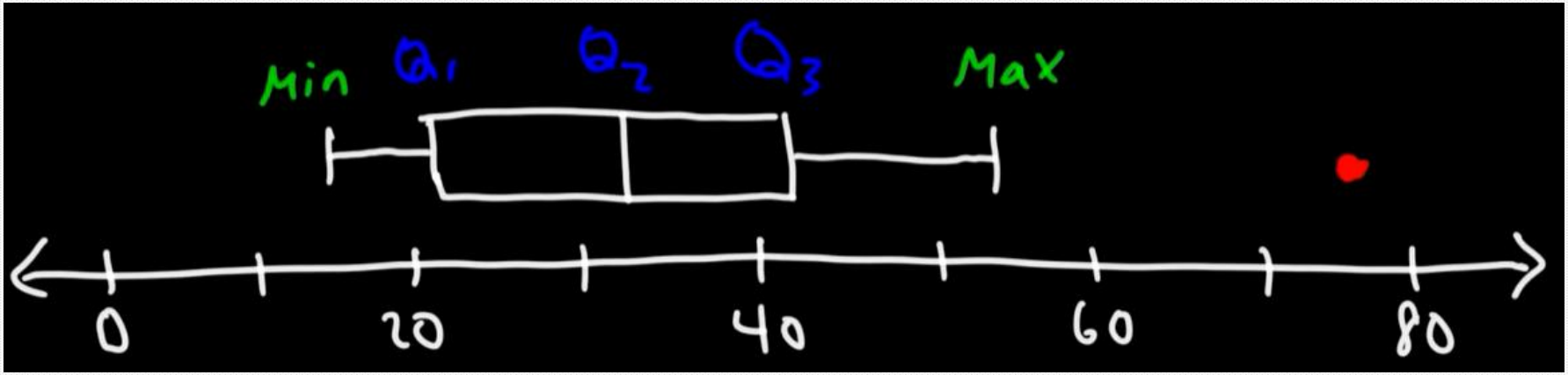
=[Q1 - 1.5 * IQR , Q3 + 1.5 * IQR]

= [22 - 1.5 * 21 , 43 + 1.5 * 21] [IQR= Q3-Q1= 43-22= 21]

= [22 -31.5 , 43 + 31.5]

= [-9.5 , 74.5]

Boxplot



Coefficient of Correlation

Coefficient of Correlation

- What is correlation
- What is coefficient of correlation
- Characteristics of coefficient of correlation
- What is linear correlation
- Types of correlation on the basis of degree of relationship

Correlation

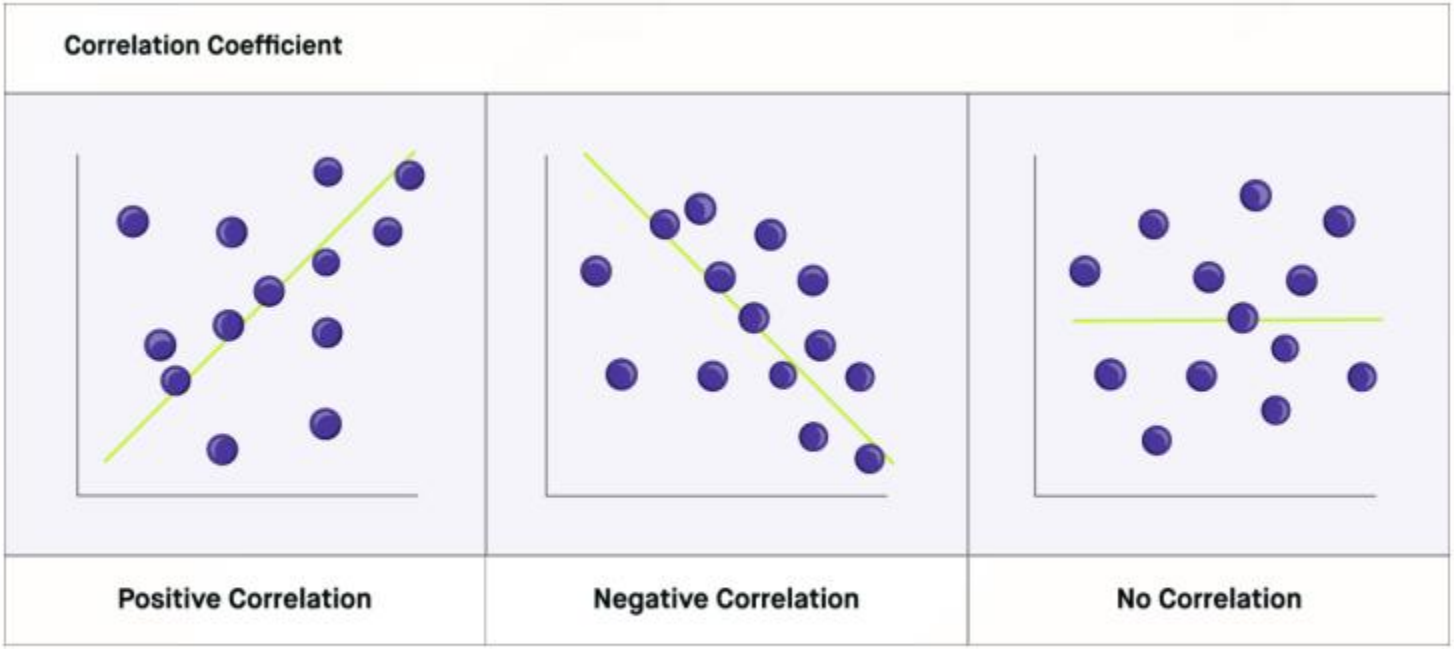
Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other

Coefficient of Correlation

A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. In other words, it reflects how similar the measurements of two or more variables are across a dataset.

Characteristics of coefficient of correlation

- It measures the relationship that exist between two variables.
- The relationship can be three types
    - A. Positive
    - B. Negative
    - C. Zero/ no/ orthogonal.
- It helps to know the direction of the relationship.
- It also helps to know the degree of relationship.
- Coefficient of correlation varies from -1 to +1.


Correlation Coefficient
Positive Correlation
Negative Correlation
No Correlation

Coefficient of correlation varies from -1 to +1

Value/Degree of Correlation	Interpretation
± 1	Perfect correlation
± 0.91 to ± 0.99	Very high correlation
±0.71 to ± 0.90	High correlation
± 0.51 to ± 0.70	Moderate correlation
± 0.31 to ± 0.50	Low correlation
± 0.11 to ± 0.30	Very low correlation
± 0.01 to ± 0.10	Almost negligible correlation
0	Zero/ no correlation

How to calculate Correlation

Age x	Glucose Level y
43	99
21	65
25	79
42	75
57	87
59	81

How to calculate Correlation

Subject	Age x	Glucose Level y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	**247**	**486**	**20485**	**11409**	**40022**

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

How to calculate Correlation

From our table:

$\Sigma x = 247$

$\Sigma y = 486$

$\Sigma xy = 20{,}485$

$\Sigma x^2 = 11{,}409$

$\Sigma y^2 = 40{,}022$

n is the sample size, in our case = 6

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

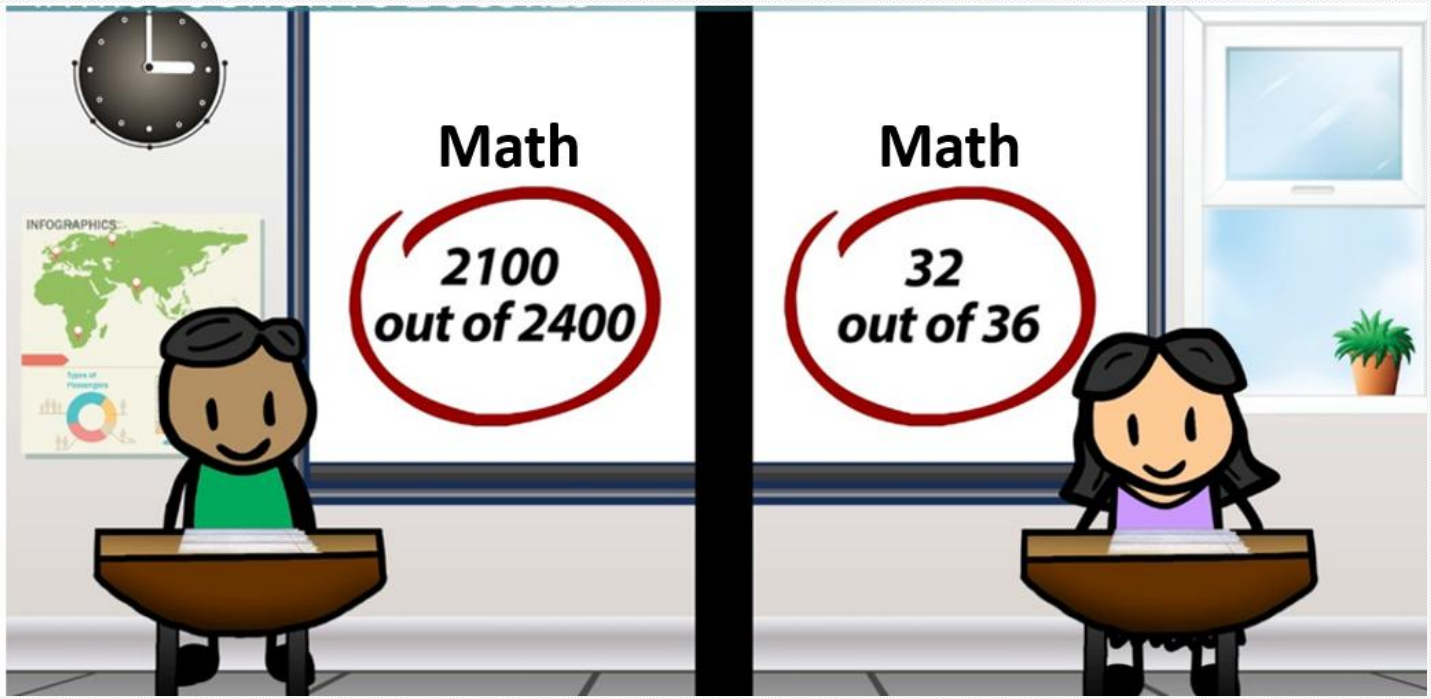
$$\text{The correlation coefficient} = \frac{6(20{,}485) - (247 \times 486)}{[\sqrt{[[6(11{,}409) - (247^2)] \times [6(40{,}022) - 486^2]]}]}$$

$$= 0.5298$$

Standard Score

(Z Score, T Score)

Standard Score



Standard Score

- Z-Score
- T-Score

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

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

	ABC Math	XYZ Math
M	70	30
SD	20	10
X	80	50
Z-Score	**0.5**	**2**

$$z = \frac{80 - 70}{20} = \frac{10}{20} = 0.5$$

$$z = \frac{50 - 30}{10} = \frac{20}{10} = 2$$

Standard Score

❑T-Score

$$T - Score = 50 + \frac{10(X - m)}{SD}$$

$$T - Score = 50 + 10Z$$

	ABC Math	XYZ Math
M	70	30
SD	20	10
X	80	50
Z-Score	**0.5**	**2**
T-Score	55	70

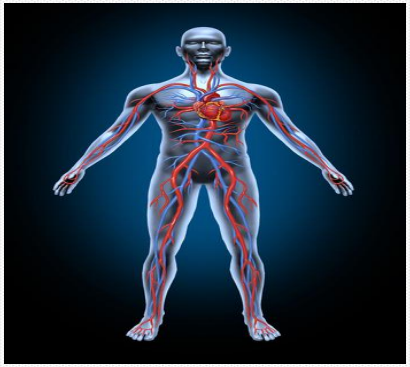
T = 50 + 10 X 0.5 = 55

T = 50 + 10 X 2 = 70

Population Data
Vs
Sample Data

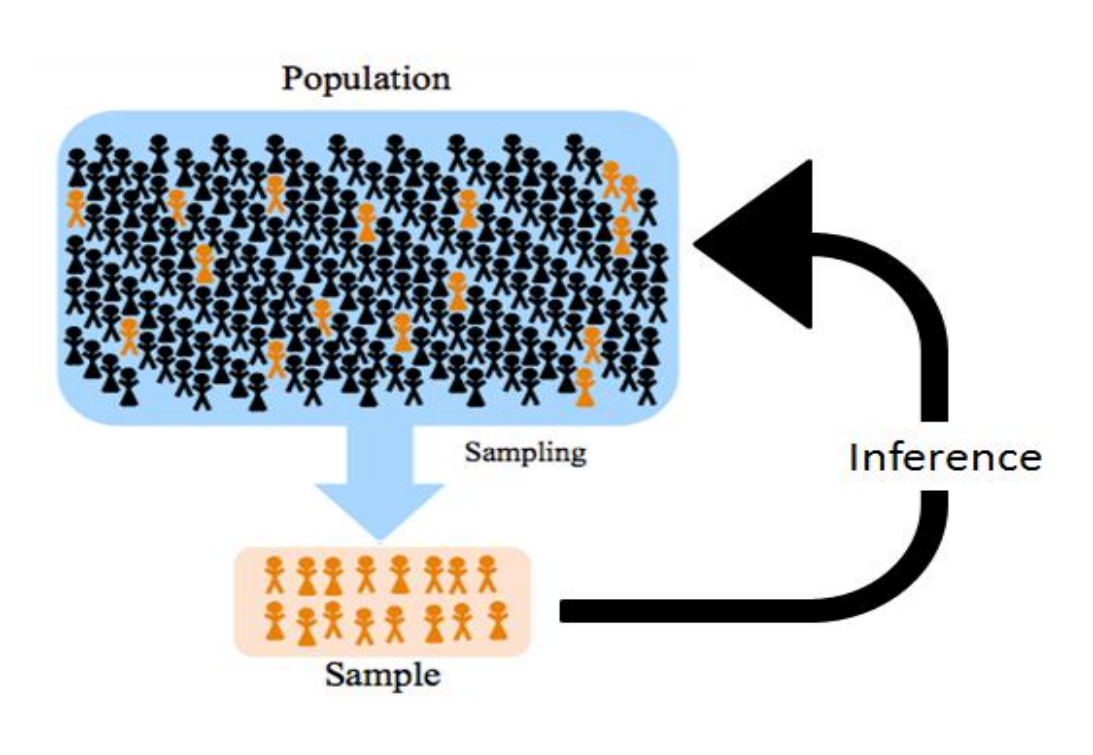
Population Vs Sample

Population Data



Sample Data

Population Vs Sample



Population Vs Sample

Population Data

A population includes all members from a specified group, all possible outcomes or measurements

Sample Data

A sample consists of some observations drawn from the population, so a part or a subset of the population.

- Sample is the representative of the Population
- Sample should be large enough to represent the Population

Statistics & Parameters

The representative values like mean, median, standard deviation, etc. calculated from the samples are called **statistics** and those directly computed from the population are named as **parameters**.

Population Vs Sample

Parameters are usually Greek letters or capital letters. Statistics are typically Roman letters or small letters.

Unit	Statistic **Sample**	Parameter **Population**
Population	p	P
Unit of element	x	X
Mean	$\bar{x}$	μ (mu)
Standard Deviation	s	σ (Sigma)
Varince	s^2	σ^2
Number of elements	n	N
Correlation Coefficient	r	ρ (rho)
Regression Coefficient	b	β (beta)

Resources Link

- Free Online Courses
  - Khan Academy – Statistics and Probability: https://www.khanacademy.org/math/statistics-probability
  - HarvardX – Statistics and R (edX): https://pll.harvard.edu/course/statistics-and-r
  - Coursera – Intro to Statistics (Audit Free): https://www.coursera.org/projects/statistics-data-science

- Free Textbooks
  - OpenIntro Statistics (Free Book): https://www.openintro.org/book/os/
  - Introductory Statistics (OpenStax): https://openstax.org/books/introductory-statistics/pages/1-introduction
  - Statistical Thinking for the 21st Century: https://statsthinking21.org/

Resources Link

- 🎥YouTube Channels
  - StatQuest with Josh Starmer:
    https://www.youtube.com/user/joshstarmer

🧠Interactive Learning

- Seeing Theory – Interactive Statistics Visualizations:
  https://seeing-theory.brown.edu/