Module 3: Statistical Inference

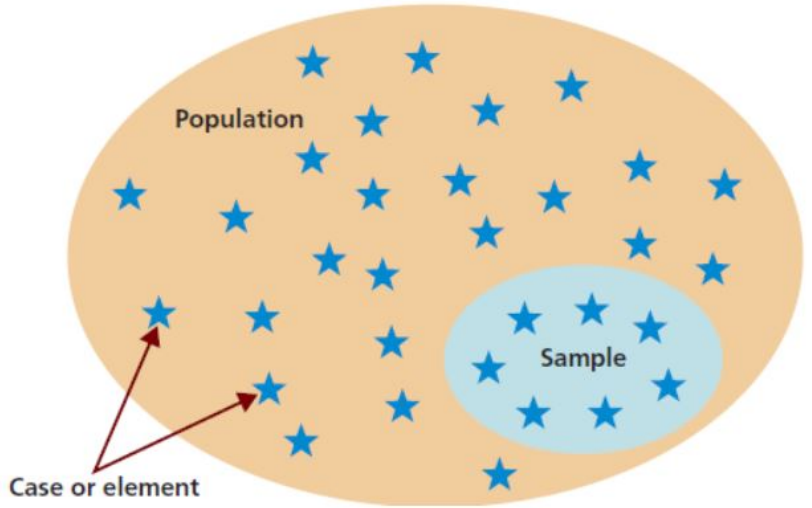
What is a Statistic?

A statistic is a piece of data from a **portion of a population**

Ex: Limitation in gathering census data for the whole population. Instead take a sample and estimate the several details of the population

Types of Statistic:

- Descriptive Statistics
- Estimators
- Test Statistics



Descriptive Statistics

- **Descriptive statistics** provide ways of capturing the properties of a given data set or sample. They summarize observed data, and provide a language to talk about it.

  Anything that describes data is **descriptive statistics.**

- There are two main types of descriptive statistics:
  - **Central tendency measures**, which capture the center around which the data is distributed.[Eg;avg. marks scored or avg.age group in the dataset]
  - **Variation or variability measures**, which describe the data spread, i.e. how far the measurements lie from the center.

Centrality Measures

- The three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

Mean

- The mean is the arithmetic average, and it is probably the measure of central tendency.
- Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.
- If you change any value, the mean changes.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

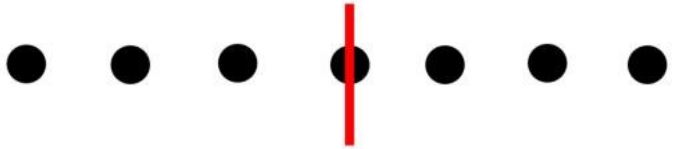
- The median is the middle value.
- It is the value that splits the dataset in half.
- To find the median, order your data from smallest to largest, and then find the data point

  that has an equal amount of values above it and below it.
- The method for locating the median varies slightly depending on whether your

  dataset has an even or odd number of values.

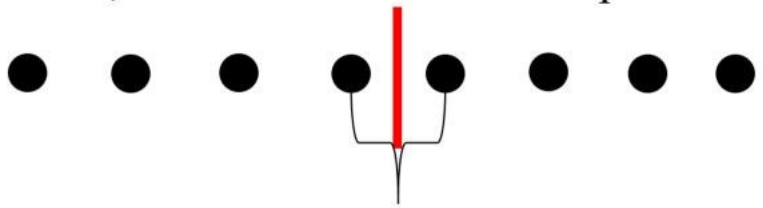
$$Location = \left(\frac{N+1}{2}\right)^{th}$$

N = 7, so 4th data point



$$Location = mid\ of\ \left(\frac{N}{2}\right)^{th}\ \&\ \left(\frac{N}{2}+1\right)^{th}$$

N = 8, so mid of 4th and 5th data point



Finding Median in a Dataset

Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

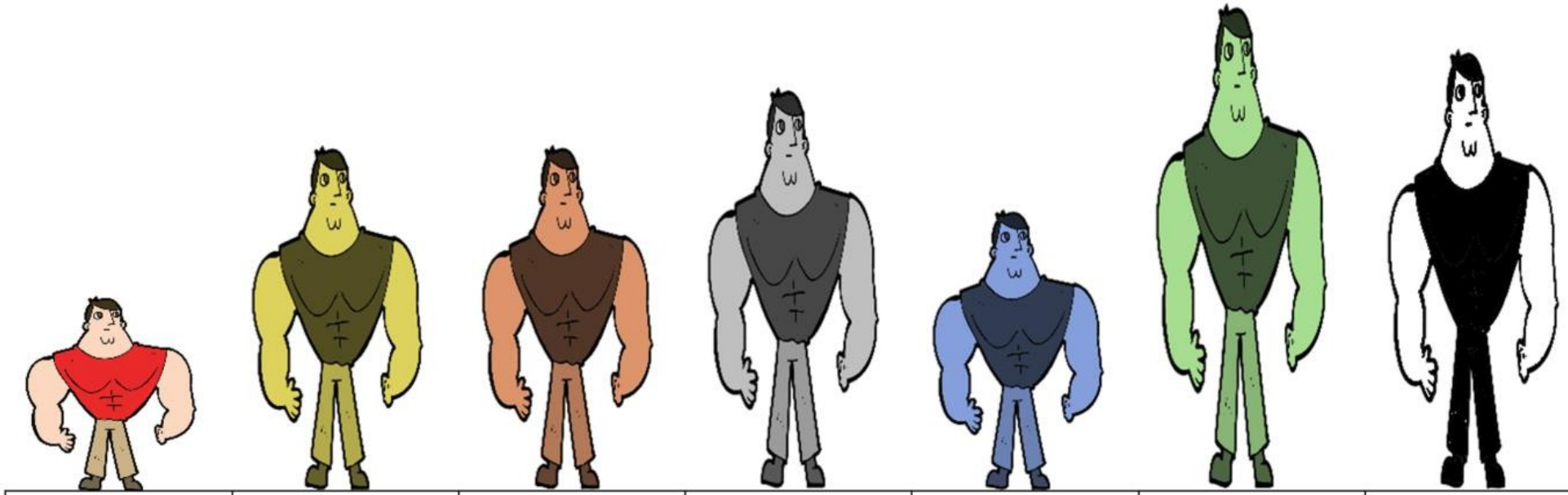
Median Even
40
38
35
33
32
30
29
28
27
26
24
23
22
19
17

Median	Median Fixed
69	112
56	93
54	89
52	82
47	47
46	46
46	46
45	45
43	43
36	36
35	35
34	34
31	31

Mode

- The mode is the value that occurs the most frequently in your data set.
- Mode is used with categorical, ordinal, and discrete data

Mode
5
5
5
4
4
3
2
2
1

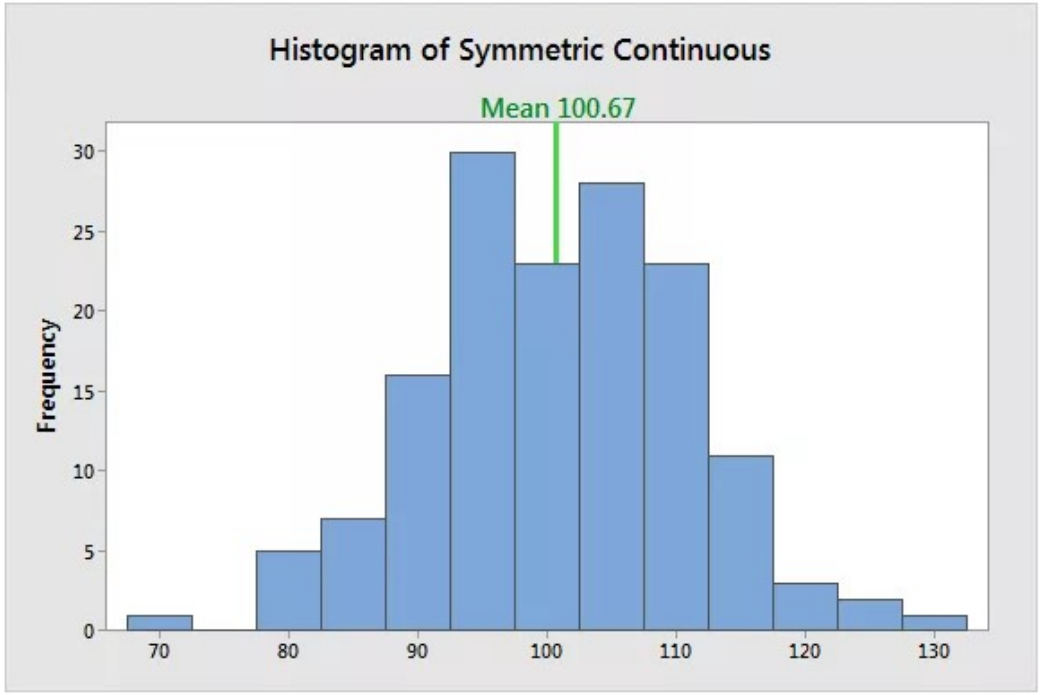


A	B	C	D	E	F	G
150 cm	160 cm	160 cm	170 cm	155 cm	180 cm	175 cm

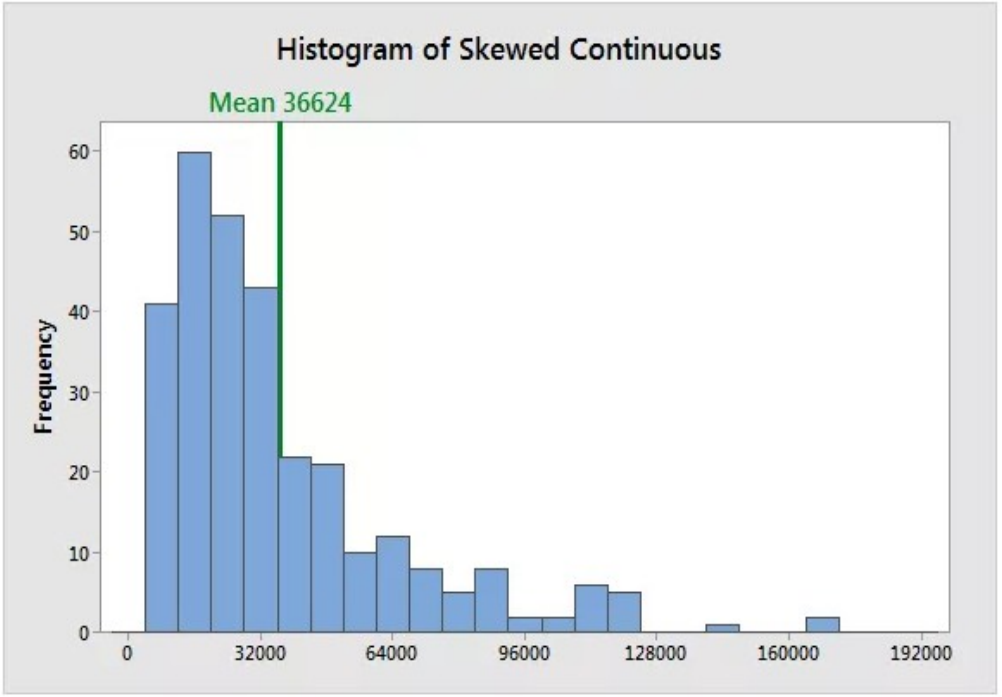
Height	150 cm	**160 cm**	170 cm	155 cm	180 cm	175 cm
Count	1	**2**	1	1	1	1

As there are two bodybuilders with 160 cm height, this implies the mode of this dataset will be 160 cm.

When to use Mean?

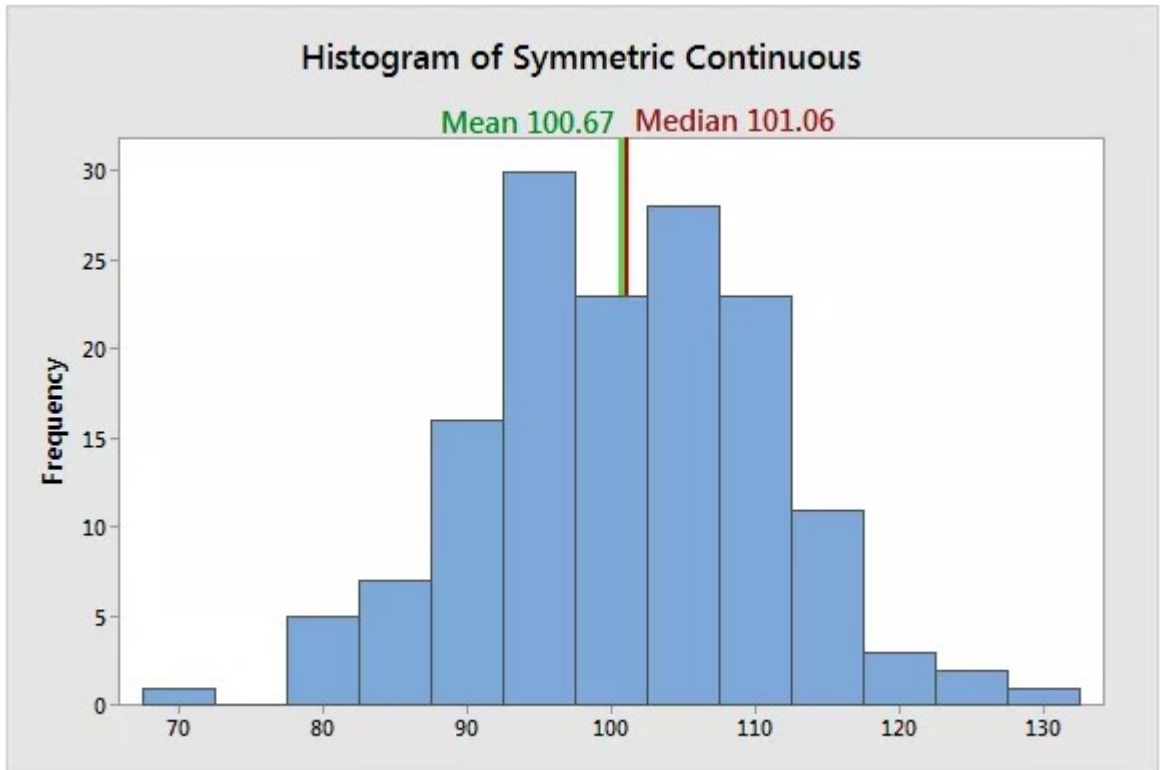


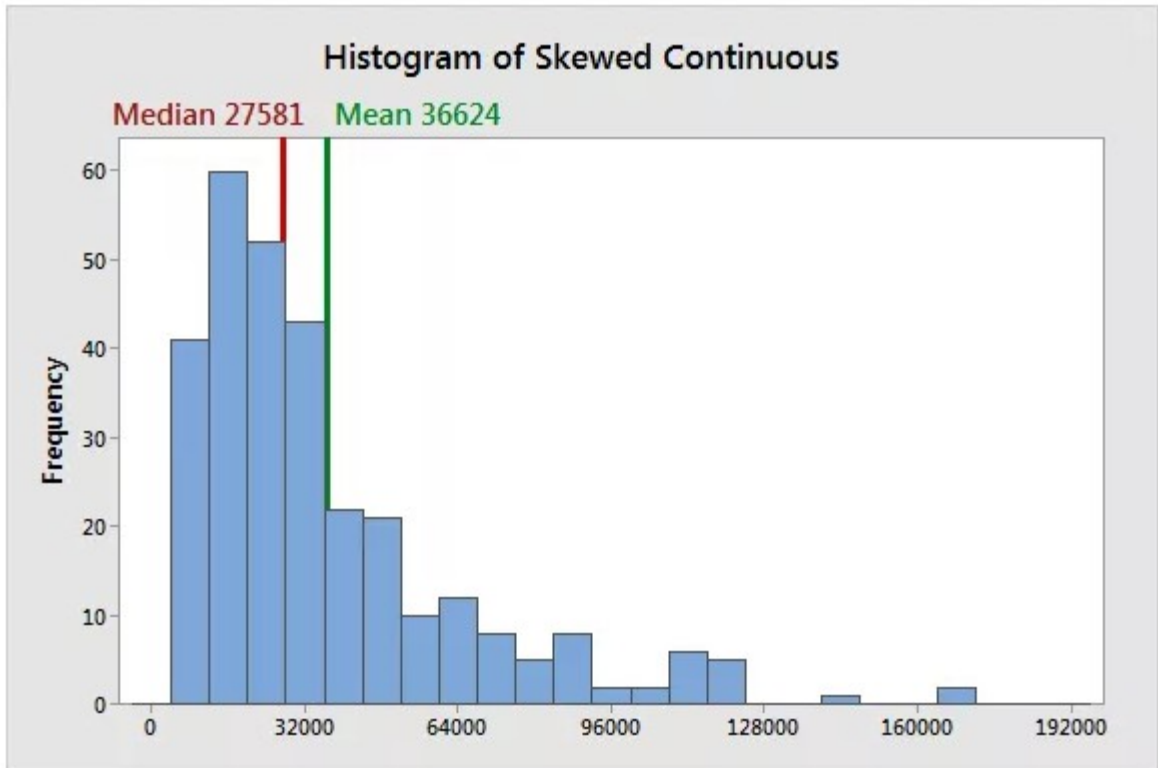
In a symmetric distribution, the mean locates the center accurately.



In a skewed distribution, the mean can miss the mark.

Comparing Mean and Median





Mean=Median ---->Normal distribution
Median <Mean ---->Positively skewed
Median >Mean ---->Negatively skewed

Variability measures

- A measure of variability is a summary statistic that represents the amount of dispersion in a dataset.
- Measures of variability define how far away the data points tend to fall from the center.

Range

- The range of a dataset is the difference between the largest and smallest values in that dataset.
- For example, in the two datasets below, dataset 1 has a range of 20 – 38 = 18

  dataset 2 has a range of 11 – 52 = 41.
- Dataset 2 has a broader range and, hence, more variability than dataset 1.

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

Standard Deviation

- The standard deviation is the standard or typical difference between each data point and the mean.
- When the values in a dataset are grouped closer together, you have a smaller standard deviation.
- On the other hand, when the values are spread out more, the standard deviation is larger because the standard distance is greater.

Variance

- Variance is the average squared difference of the values from the mean.
- The variance includes all values in the calculation by comparing each value to the mean.
- To calculate this statistic, you calculate a set of squared differences between the data points and the mean, sum them, and then divide by the number of observations. Hence, it's the average squared difference.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

	x	x-x`	(x-x`)^2
1	15	0	0
2	15	0	0
3	15	0	0
4	14	-1	1
5	16	1	1
Σ	75		2
Mean	**(75/5)=15**	**Variance**	**(2/5)=0.4**
		SD	**√0.4 =0.632456**

values are grouped close

	x	x-x`	(x-x`)^2
1	2	-13	169
2	7	-8	64
3	14	-1	1
4	22	7	49
5	30	15	225
Σ	75		508
Mean	**(75/5)=15**	**Variance**	**(508/5)=101.6**
		SD	**√101.6 =10.07968**

values are spread out

Problem

1,2,3,4,5,1,2,3,1,2,4,5,2,3,1,1,2,3,5,6

Problem

1,2,3,4,5,1,2,3,1,2,4,5,2,3,1,1,2,3,5,6

Mean = 2.8

Median = 2.5

Mode = 1, 2

Variance = 2.48

standard Deviation = 1.567

Problem

1. A cricketer's scores in five ODI matches are as follows: 12, 34, 45, 50, 24. Find the mean, Median and Range of the scores.
2. Find the mean, Range and median of 4, 4, 6, 3, 2.
3. Find the variance and Standard deviation of 50, 67, 24, 34, 78, 43.
4. Find the mean, median and mode of 6, 8, 9, 3, 4, 6, 7, 6, 3.
5. Find the mean of the data which has mode = 65 and median = 61.6. (2Mean + Mode = 3Median)
6. Tabulate the difference between mean and median.
7. Find the mean, median, mode, range, standard deviation and variance of the following data:

(a) 9, 7, 11, 13, 2, 4, 5, 5

(b) 16, 18, 19, 21, 23, 23, 27, 29, 29, 35

(c) 2.2, 10.2, 14.7, 5.9, 4.9, 11.1, 10.5

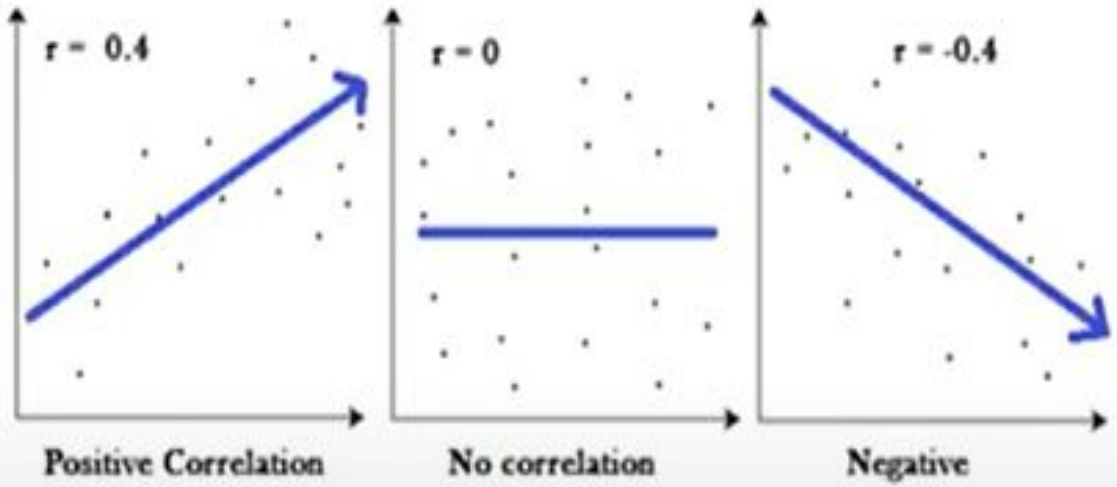
Correlation in Datascience

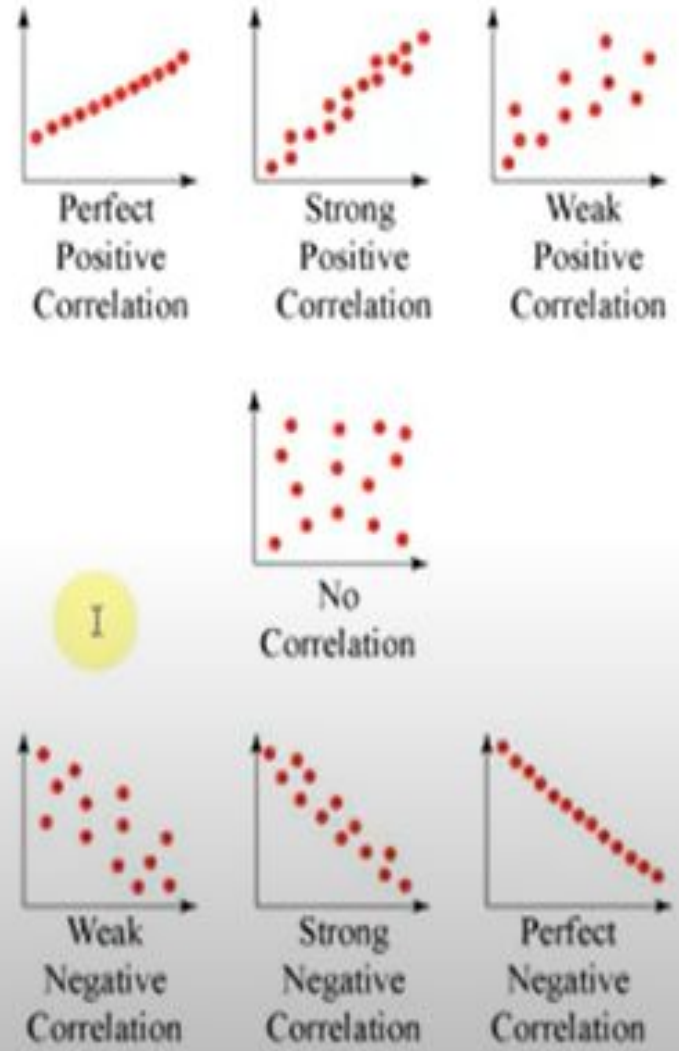
Correlation

- Correlation (to be exact Correlation in Statistic) is a measure of a mutual relationship between two variables whether they are causal or not.
- Correlation is useful because it can indicate a predictive relationship that could be exploited in the practice.
- For example, the number of Ice cream sold are more frequently with a higher temperature in the day.

Correlation coefficients are used to measure how strong a relationship is between two variables.

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.





Types of correlation coefficient formulas.

Covariance and Correlation

- A subset of the population is called a sample.
- Correlation and covariance are calculated on samples and not populations termed as sample covariance and correlation.
- Both terms define the relationship and dependency between the variables.

- Correlation measures the association between the variables.

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation

Correlation between X and Y

Standard deviation of X

Standard deviation of Y

Covariance explains the joint variability of the variables.

$$\text{Cov}(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Where

xi = data value of x
yi = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values.

Pearson Correlation Coefficient

- Pearson Correlation is one of the most used correlations during the data analysis process.
- Pearson correlation measures the linear relationship between variable continuous X and variable continuous Y and has a value between 1 and -1.
- In other words, the Pearson Correlation Coefficient measures the relationship between 2 variables via a line.

Pearson Correlation Coefficient

- When the correlation coefficient is **closer to value 1**, it means there is a **positive relationship** between variable X and Y. A positive relationship indicates an increase in one variable associated with an increase in the other.
- On the other hand, **the closer correlation coefficient is to -1** would mean there is a **negative relationship** which is the increase in one variable would result in a decrease in the other.
- **If *X* and *Y* are independent**, then the **correlation coefficient is close to 0** although the Pearson correlation can be small even if there is a strong relationship between two variables.

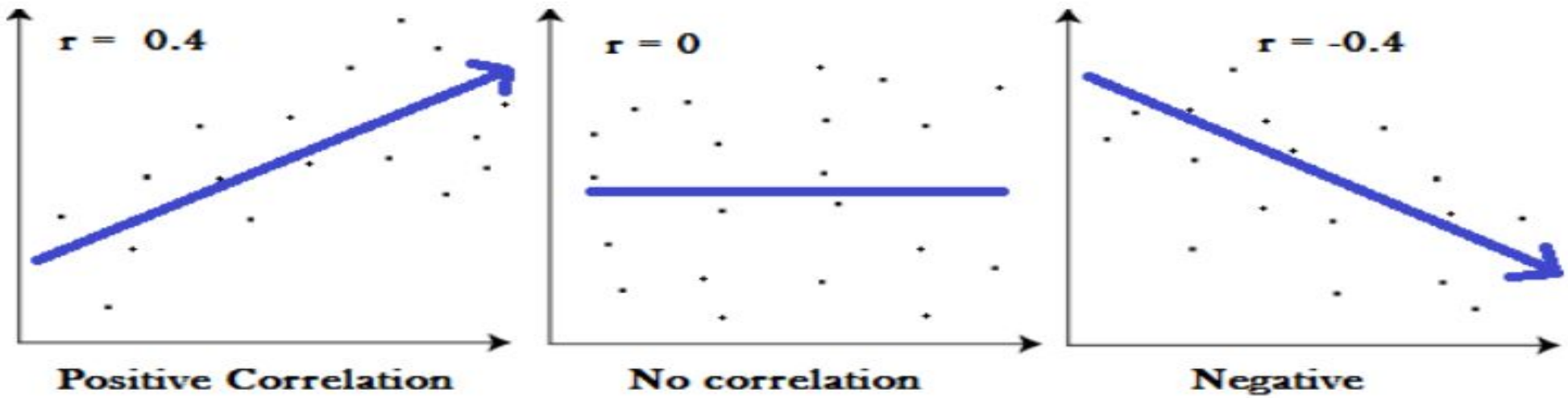
Pearson Correlation Coefficient Formula

- Where,
- r = Pearson Coefficient
- n= number of the pairs of the stock
- $\sum xy$ = sum of products of the paired stocks
- $\sum x$ = sum of the x scores
- $\sum y$= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

- A Pearson correlation coefficient of between 0 and 0.3 (or 0 and -.03) indicates a weak relationship between the two variables

- A Pearson correlation coefficient of between 0.4 and 0.6 (or -.04 and -.06) indicates a moderate strength relationship between the two variables

- A Pearson correlation coefficient of between 0.7 and 1 (or -.07 and 1) indicates a strong relationship between the two variables.



Subject	Age x	Glucose Level y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

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

III IA skill based

- Skill for all
- Project title

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

The answer is: **2868 / 5413.27 = 0.529809**

Click here if you want easy, step-by-step instructions for solving this formula.

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

The correlation coefficient =

- $6(20{,}485) - (247 \times 486) / [\sqrt{[[6(11{,}409) - (247^2)] \times [6(40{,}022) - 486^2]]}$
  $= 0.5298$

Spearman's rank correlation

- Pearson's correlation captures correlations of first order, but not nonlinear correlations.
- It does not work well in the presence of outliers.
- Solution : The Spearman's rank correlation when the data contain outliers.
- The main idea is to use the ranks of the sorted sample data, instead of the values themselves
- Spearman's correlation computes the correlation between the ranks of the data.

Ex:

- Data -[4, 3, 7, 5]
- Ordered list ([3, 4, 5, 7])
- Rank of 4 is 2

Spearman Rank Correlation

Unlike the Pearson Correlation Coefficient, Spearman Rank Correlation measures the monotonic relationship (Strictly increase or decrease, not both) between two variables and measured by the rank order of the values.

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

monotonic function is one that either never increases or never decreases as its independent variable changes.

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

We then complete the following table:

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

as n = 10. Hence, we have a ρ (or r_s) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Problem 2

Sales	Ad
90	7
85	6
68	2
75	3
82	4
80	5
95	8
70	1

Sales	Ad	Ran for Sales (x)	Rank for Ad (y)	d = x - y	d^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	-1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	-1	1
Total	-	-	-	0	4

Since n = 8 and ∑d2 = 4, apply the above formula, we get

r = 1 - 6∑d2/n(n2 - 1)

= 1 - 6x4/8(82 - 1)

Since n = 8 and ∑d2 = 4, apply the above formula, we get

r = 1 - 6∑d2/n(n2 - 1)

= 1 - 6x4/8(82 - 1)

= 1 - 0.0476

= 0.95

The high positive value of the rank correlation coefficient indicates that there is a very good amount of agreement between sales and advertisement.

Repeated ranks

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

where m_i is the number of repetitions of i^{th} rank

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{[81.5 + 0.5 + 2]}{504} = 1 - \frac{504}{504} = 0$$

Marks in Commerce and Mathematics are uncorrelated

- A Spearman's correlation coefficient of between 0 and 0.3 (or 0 and -.03) indicates a weak monotonic relationship between the two variables

- A Spearman's correlation coefficient of between 0.4 and 0.6 (or -.04 and -.06) indicates a moderate strength monotonic relationship between the two variables

- A Spearman's correlation coefficient of between 0.7 and 1 (or -.07 and 1) indicates a strong monotonic relationship between the two variables.

What is monotonicity?

As per the OXFORD dictionary, Monotonic is a function or quantity varying in such a way that it either never decreases or never increases.

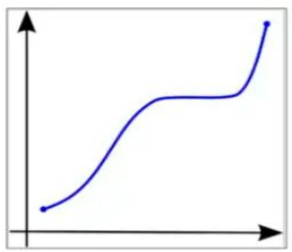


Figure 1 - A monotonically increasing function

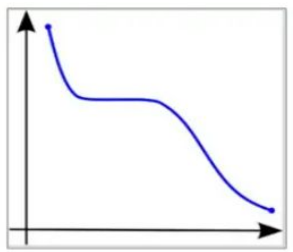


Figure 2 - A monotonically decreasing function

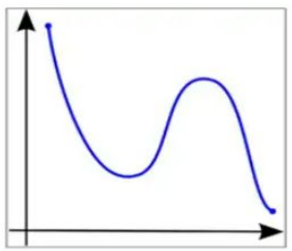
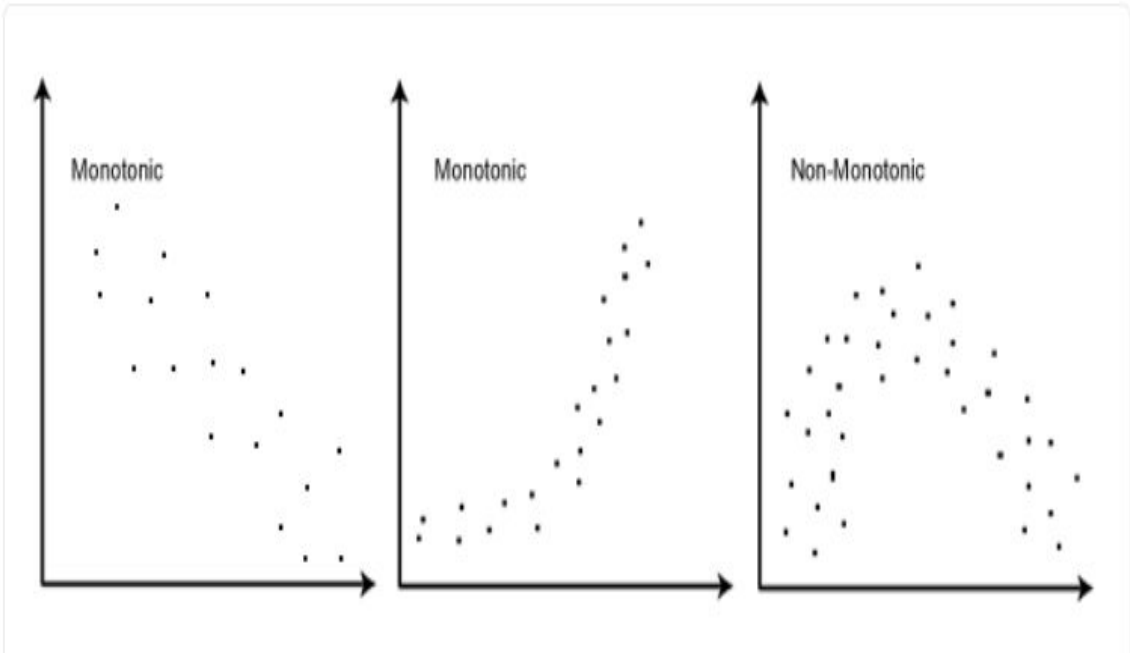


Figure 3 - A function that is not monotonic

Source: https://en.wikipedia.org/wiki/Monotonic_function

What is a monotonic relationship?

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. Examples of monotonic and non-monotonic relationships are presented in the diagram below:



Pearson Vs. Spearman correlation methods

Confused about when to use the Pearson correlation and when to use the Spearman's correlation coefficient? Remember that Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines. Linear relationships are straight line relationships. Monotonic relationships differ from linear relationships in that the two variables might converge, but not at a constant rate. There are three types of monotonic functions:

- **Monotonically increasing relationships**

This means that as the x variable increases, the y variable never decreases.

- **Monotonically decreasing relationships**

This means that as the x variable increases, the y variable never increases.

- **Non monotonic relationships**

This means that as the x variable increases, the y variable sometimes decreases and sometimes increases.

Comparison of Pearson and Spearman coefficients

--The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.

-- Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.

Kendall Tau Rank Correlation

- Kendall Tau rank correlation coefficient measures the degree of similarity between two sets of ranks given to the same set of objects. This coefficient is more appropriate for discrete data.

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$