Statistics Terminologies

Population Vs Sample

Statistics

The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

Population

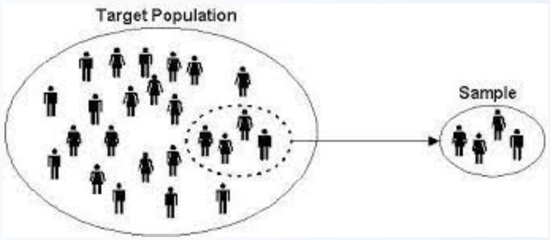
Generally, population refers to the people who live in a particular area at a specific time.

But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc.

Sample

A sample is defined as a smaller and more manageable representation of a larger group.

A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.



Sampling

Simple Random Sampling

Each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected

Sampling

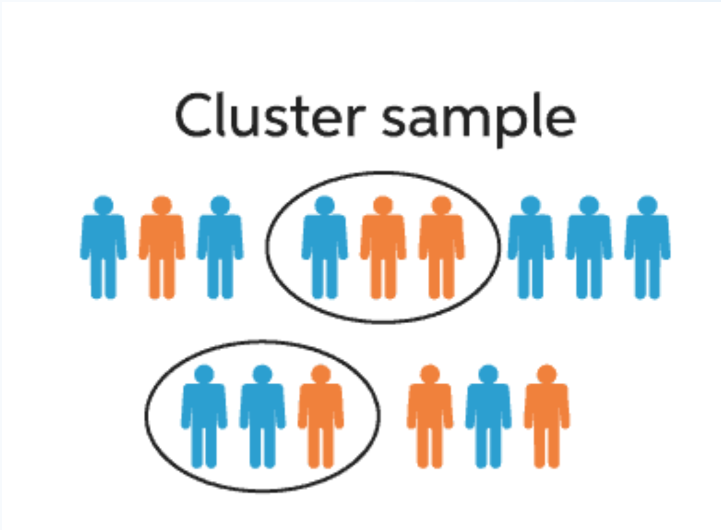
Stratified Sampling:

The population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups

Example—A student council surveys 100 students by getting random samples of 25 freshmen, 25 sophomores, 25 juniors, and 25 seniors.

Sampling

Cluster random sample: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.



Example—An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.

Sampling

Systematic random sample: Members of the population are put in some order. A starting point is selected at random, and every nth member is selected to be in the sample.

Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 20th student is selected to take a survey.

Parameter vs Statistic

Parameter

Parameters are numbers that describe the properties of entire populations.

Statistic

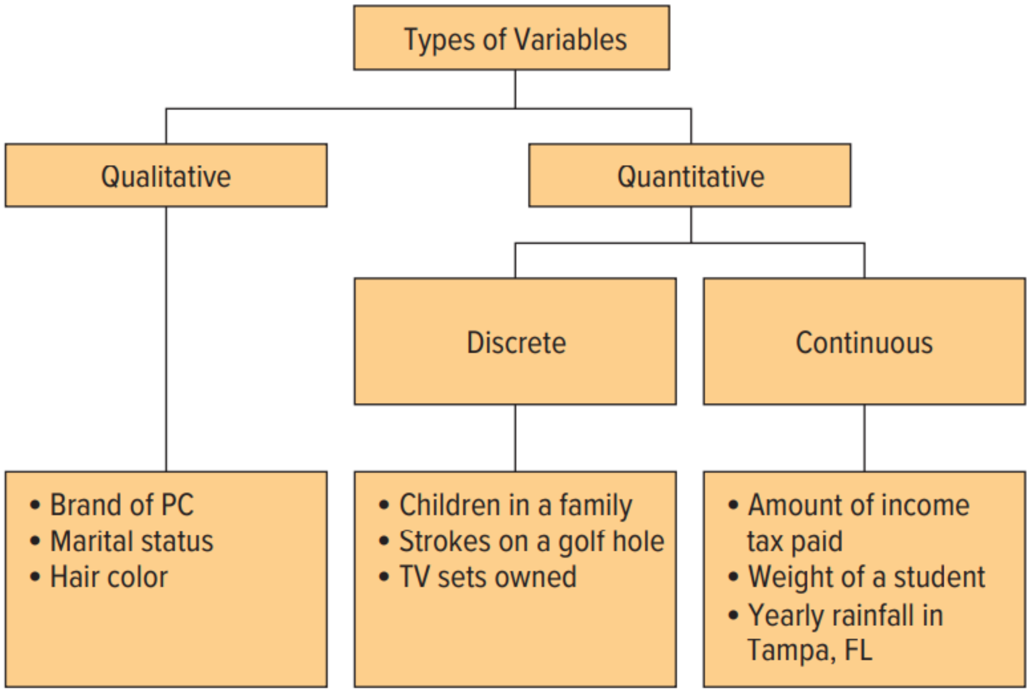
Statistics are numbers that describe the properties of samples.

Example :

The average income for the India is a population parameter.

The average income for a sample drawn from the India is a sample statistic.

Types of Variables



Types of Variables

- **Categorical Variable**: variables than can be put into categories. For example, the category "Toothpaste Brands" might contain the variables Colgate and Aquafresh.

- **Dependent Variable** : the outcome of an experiment. As you change the independent variable, you watch what happens to the dependent variable.
- **Independent Variable**: a variable that is not affected by anything that you, the researcher, does. Usually plotted on the x-axis.

- **Discrete Variable**: a variable that can only take on a certain number of values. For example, "number of cars in a parking lot" is discrete because a car park can only hold so many cars.
- **Continuous variable**: a variable with infinite number of values, like "time" or "weight".

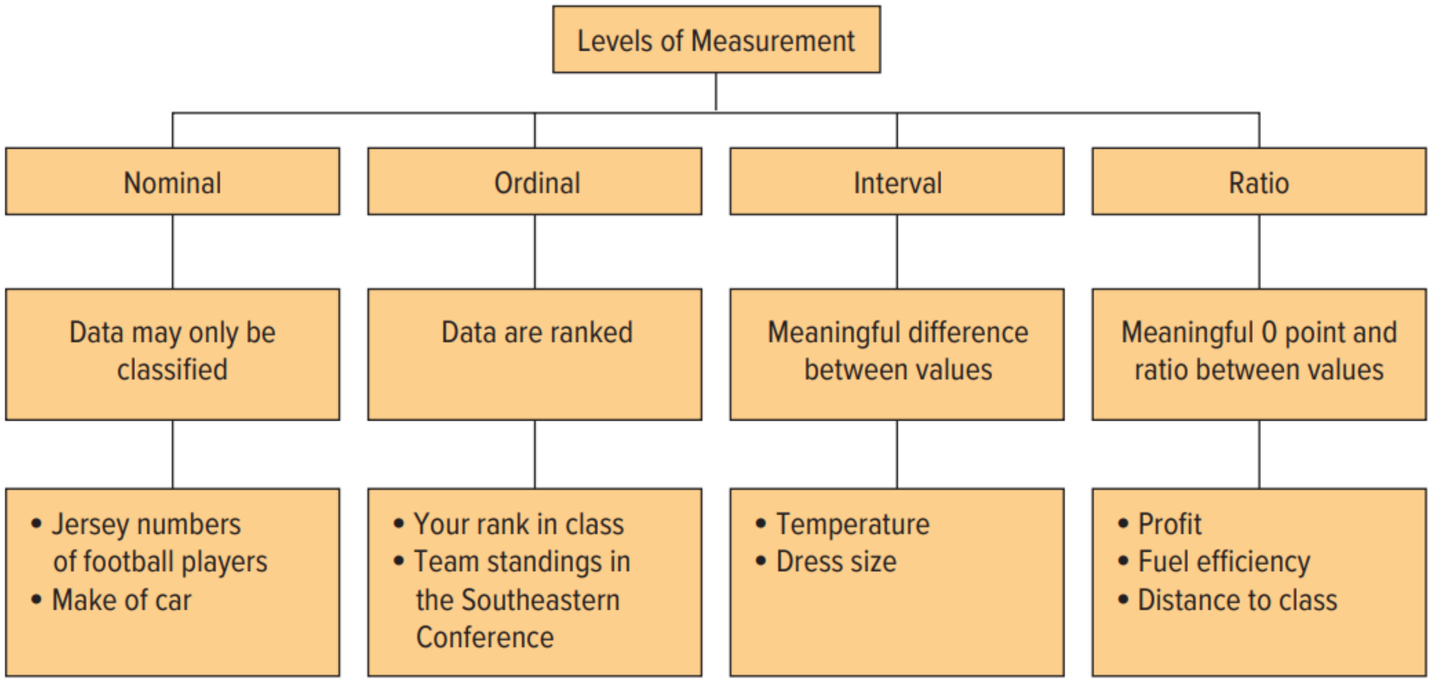
Types of Variables

- **Nominal Variable**: another name for categorical variable.
- **Ordinal Variable**: similar to a categorical variable, but there is a clear order. For example, income levels of low, middle, and high could be considered ordinal.

- **Qualitative Variable**: a broad category for any variable that can't be counted (i.e. has no numerical value). Nominal and ordinal variables fall under this umbrella term.
- **Quantitative Variable:** A broad category that includes any variable that can be counted, or has a numerical value associated with it. Examples of variables that fall into this category include discrete variables and ratio variables.

- **Random Variable:** are associated with random processes and give numbers to outcomes of random events.
- **Ranked Variable :** is an ordinal variable; a variable where every data point can be put in order (1st, 2nd, 3rd, etc.).

Levels of Measurement



Descriptive Statistics : Central Tendency

Central Tendency

Measures of Central Tendency

Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution. It represents the single value of the entire population or a dataset.

We will consider five Measures of Central Tendency

- **the arithmetic mean**
- **the median**
- **the mode**
- **the weighted mean**
- **the geometric mean.**

Arithmetic Mean

Population Mean

the population mean is the sum of all the values in the population divided by the number of values in the population. To find the population mean, we use the following formula.

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

It is denoted by μ

$$\mu = \frac{\Sigma x}{N}$$

Example

Listed below are the distances between exits (in miles).

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

$$\mu = \frac{\Sigma x}{N} = \frac{11 + 4 + 10 + \cdots + 1}{42} = \frac{192}{42} = 4.57$$

Arithmetic Mean

Properties of Arithmetic mean:

The arithmetic mean is a widely used measure. It has several important properties:

1) To compute a mean, the data must be measured at the interval or ratio level.
2) The mean is unique. That is, there is only one mean in a set of data. we will discover a measure of central tendency that may have more than one value.

Disadvantage of Arithmetic mean

- One of the major drawbacks of arithmetic mean is that it is changed by extreme values in the data set.
- It is not an appropriate average for highly skewed distributions.
- It cannot be computed accurately if any item is missing.

Median

Median

When our data contains one or two very large or very small values, the arithmetic mean may not be representative. The center for such data is better described by a measure of location called the median.

The midpoint of the values after they have been ordered from the minimum to the maximum values.

Formula for median

n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$$

Where n is total number of data points in our sample.

Median

Example 1 (Odd numbers):

21,32,65,40,30,90,26

Ordered list : 21,26,30,32,40,65,90

n = 7

Median

= $(\frac{\mathbf{7+1}}{\mathbf{2}})^{th}$ Observation

= 4^{th} observation = 32

Example 1 (Even numbers):

10,9,7,12,8,11

Ordered list : 7,8,9,10,11,12

n = 6

Median

$$= \frac{(\frac{6}{2})^{th}\ observation + (\frac{6}{2}+1)^{th} observation}{2}$$

$$= \frac{3^{rd}\ observation + 4^{th}\ observation}{2}$$

$$= \frac{9+10}{2}$$

= 9.5

Mode

In a dataset mode is the value of the observation that **appears most frequently.**
The value of the observation that appears most frequently**.**

Example :

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

The frequency table for above data is shown below

Distance in Miles between Exits	Frequency
1	8
2	7
3	7
4	3
5	4
6	1
7	3
8	2
9	1
10	4
11	1
14	1
Total	42

As we can see that value 1 is occurring most number of time (8 times) the mode is 1.

Note* : A dataset can consist of more than 1 mode which is called multimodal dataset.

Weighted mean

The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value.

the weighted mean of a set of numbers designated x1, x2, x3, . . . , xn with the corresponding weights w1, w2, w3, . . . , wn is computed by:

WEIGHTED MEAN

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

EXAMPLE

The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees, 14 of whom are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid to the 26 employees?

SOLUTION

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula (3–3), the mean hourly rate is:

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The weighted mean hourly wage is rounded to \$18.12.

Geometric mean

The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time.
The geometric mean of a set of n positive numbers is defined as the nth root of the product of n values. The formula for the geometric mean is written:

GEOMETRIC MEAN $$GM = \sqrt[n]{(x_1)(x_2)\cdots(x_n)}$$

As an example of the geometric mean, suppose you receive a 5% increase in salary this year and a 15% increase next year. The average annual percent increase is 9.886%, not 10.0%. Why is this so? We begin by calculating the geometric mean. Recall, for example, that a 5% increase in salary is 105%. We will write it as 1.05.

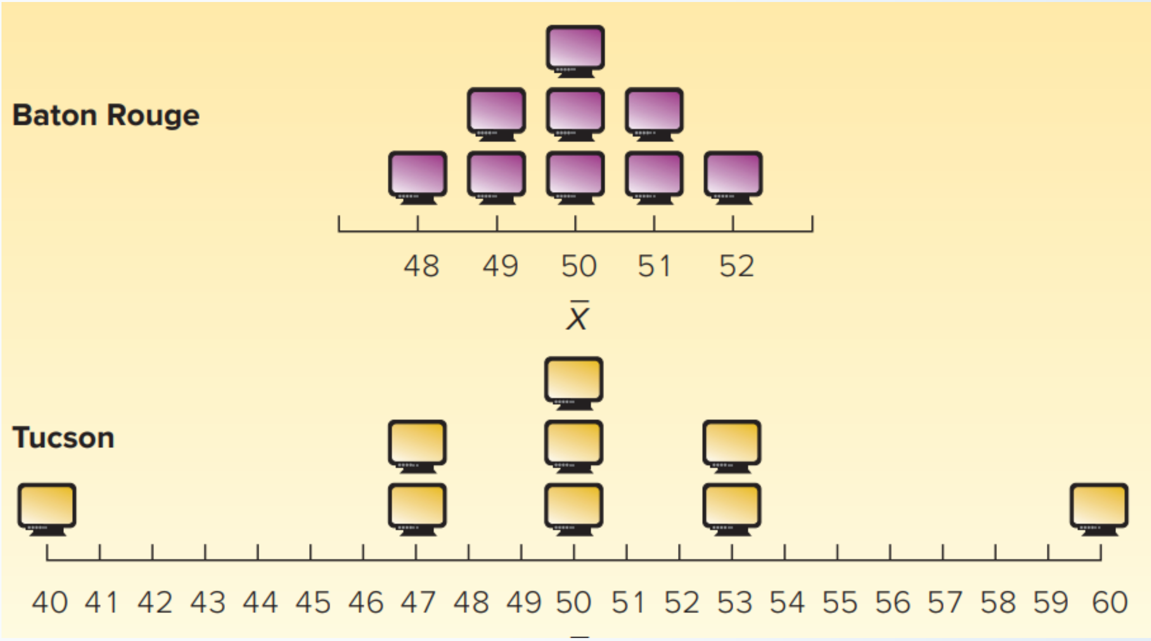
$$GM = \sqrt{(1.05)(1.15)} = 1.09886$$

Measures of Dispersion

Measure of Dispersion

Why should we study Measure of Dispersion?

A measure of location, such as the mean, median, or mode, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data.



Range

The simplest measure of dispersion is the range.
It is the difference between the maximum and minimum values in a data set.

Range = Maximum value − Minimum value

Example :

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

In the above dataset
Minimum = 1
Maximum = 14
Therefore
Range = 14-1 = 13

Variance

A limitation of the range is that it is based on only two values, the maximum and the minimum; it does not take into consideration all of the values. The variance does.

It measures the mean amount by which the values in a population, or sample, vary from their mean.

VARIANCE The arithmetic mean of the squared deviations from the mean.

Formula for variance is

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

Example :

	A	B	C
1		California Airports	
2		Orange County	Ontario
3		20	20
4		40	45
5		50	50
6		60	55
7		80	80
8			
9	Mean	50	50
10	Median	50	50
11	Range	60	60

F	G	H
Calculation of Variance for Orange County		
Number Sold	Each Value - Mean	Squared Deviation
20	20 - 50 = -30	900
40	40 - 50 = -10	100
50	50 - 50 = 0	0
60	60 - 50 = 10	100
80	80 - 50 = 30	900
	Total	2000

Source: Microsoft Excel

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{N} = \frac{(-30^2) + (-10^2) + 0^2 + 10^2 + 30^2}{5} = \frac{2{,}000}{5} = 400$$

Population Standard deviation

When we compute the variance, it is important to understand the unit of measure and what happens when the differences in the numerator are squared.

Units of standard deviation is same as units of our dataset.

Standard deviation is squared root of variance

Formula:

STANDARD DEVIATION $$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Sample Variance and Sample Standard deviation

Sample Variance

The formula for the sample variance is:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

where:

- s^2 is the sample variance.
- x is the value of each observation in the sample.
- $\bar{x}$ is the mean of the sample.
- n is the number of observations in the sample.

Sample Standard deviation

Sample standard deviation is square root of sample variance and it is denoted by s.

Why n-1 instead of n?

Although the use of n is logical since $\bar{x}$ is used to estimate μ, it tends to underestimate the population variance, σ^2.

The use of (n − 1) in the denominator provides the appropriate correction for this tendency.

Because the primary use of sample statistics like s2 is to estimate population parameters like σ2, (n − 1) is used instead of n in defining the sample variance

Measures of Position

Quartiles, Deciles and Percentiles

Measures of position

The standard deviation is the most widely used measure of dispersion. However, there are other ways of describing the variation or spread in a set of data.
One method is to determine the location of values that divide a set of observations into equal parts.
These measures include quartiles, deciles, and percentiles.

Quartiles

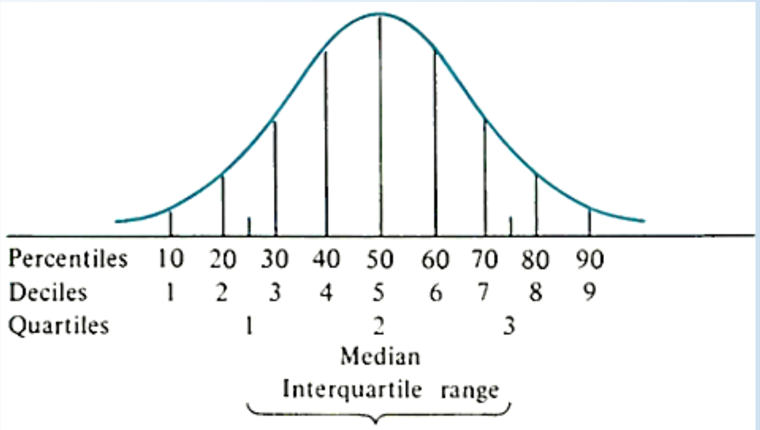
- Quartiles divide a set of observations into four equal parts.
- The first quartile, usually labelled Q1, is the value below which 25% of the observations occur
- The third quartile, usually labelled Q3, is the value below which 75% of the observations occur.
- The second quartile, usually labelled as Q2 or **Median** is the value below which 50% and above which 50% of the observations occur.

Deciles

Deciles divide a set of observations into 10 equal parts

Percentiles

Percentiles divide a set of observations into 100 equal parts



Skewness

- Skewness is a measure of the asymmetry of a distribution.
- A distribution is asymmetrical when its left and right side are not mirror images.
- A distribution can have right (or positive), left (or negative), or zero skewness.

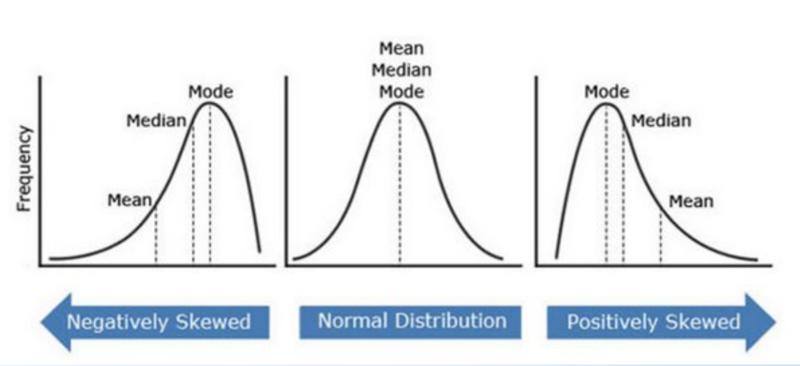
Central moments	Raw data	Discrete data	Continuous data $d' = \frac{(x-\bar{x})}{c}$
μ_1	$\frac{\sum(x-\bar{x})}{n}=0$	$\frac{\sum f(x-\bar{x})}{N}=0$	$\frac{\sum fd'}{N}\times c$
μ_2	$\frac{\sum f(x-\bar{x})^2}{N}$	$\frac{\sum f(x-\bar{x})^2}{N}=\sigma^2$	$\frac{\sum fd'^2}{N}\times c^2$
μ_3	$\frac{\sum(x-\bar{x})^3}{n}$	$\frac{\sum f(x-\bar{x})^3}{N}$	$\frac{\sum fd'^3}{N}\times c^3$
μ_4	$\frac{\sum(x-\bar{x})^4}{n}$	$\frac{\sum f(x-\bar{x})^4}{N}$	$\frac{\sum fd'^4}{N}\times c^4$

Value	Interpretation
-3 to 0	Negative Skewness
0	No Skewness
0 to 3	Positive Skewness

Karl Pearson's Coefficient of Skewness $= \frac{Mean - Mode}{Standard\ Deviation}$ or $\frac{3(Mean - Median)}{Standard\ Deviation}$

Moment based measure of skewness $= \beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}$

Pearson's coefficient of skewness $= \gamma_1 = \sqrt{\beta_1}$



Kurtosis

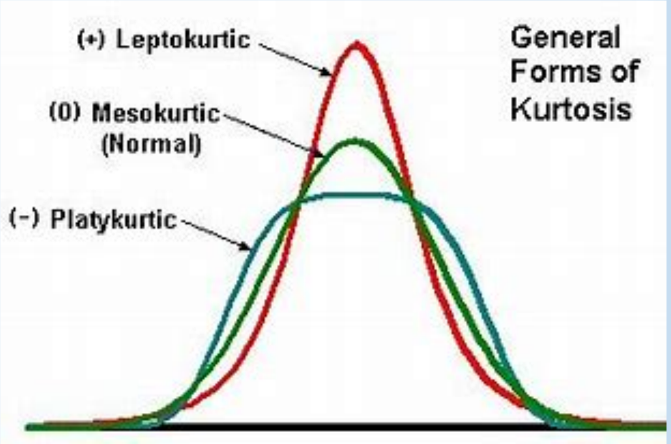
Kurtosis refers to the degree of peak of a frequency curve.
It tells how tall and sharp the central peak is, relative to a standard bell curve of a distribution.

Formula for Kurtosis

Kurtosis is measured in the following ways:

Moment based Measure of kurtosis = $\beta_2 = \dfrac{\mu_4}{\mu_2^{\,2}}$

Coefficient of kurtosis = $\gamma_2 = \beta_2 - 3$



Kurtosis can be described in the following ways:

- **Platykurtic**– When the kurtosis < 0, the frequencies throughout the curve are closer to be equal (i.e., the curve is more flat and wide)
- **Leptokurtic**– When the kurtosis > 0, there are high frequencies in only a small part of the curve (i.e., the curve is more peaked)
- **Mesokurtic**- When the kurtosis = 0 To show the peakedness of a distribution

- Platykurtic: flat and spread out
- Leptokurtic: high and thin
- Mesokurtic: normal in shape

Covariance

Covariance

Covariance measures the direction of the relationship between two variables.
A positive covariance means that both variables tend to be high or low at the same time.
A negative covariance means that when one variable is high, the other tends to be low.
A covariance of zero indicates that there is no clear directional relationship between the variables being measured.

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Note : The covariance can range from negative infinity to positive infinity.Thus, the value for a perfect linear relationship depends on the data.

Correlation

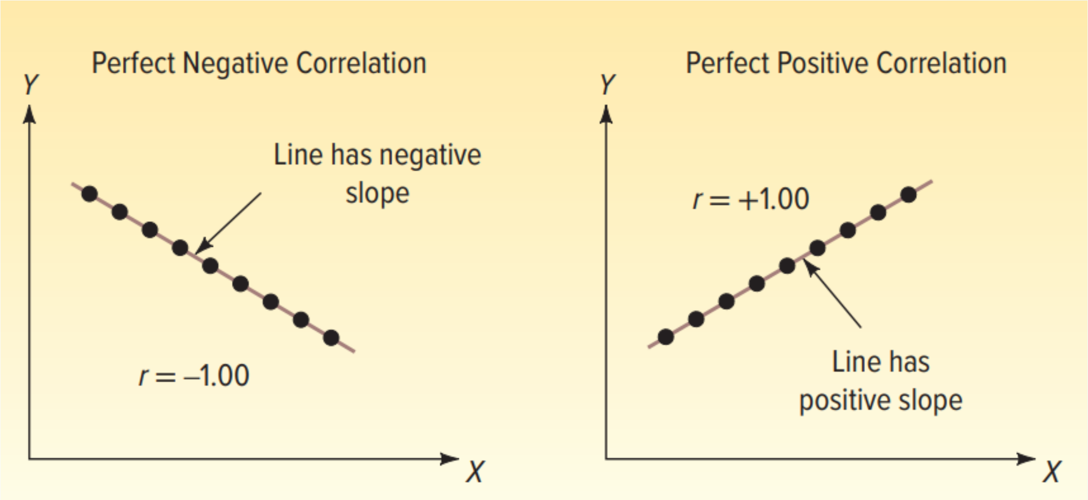
Pearson's Correlation

Pearson's correlation coefficient can be calculated to measure the direction and strength of the relationship between two variables.

The correlation coefficient is computed as:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

The sample correlation coefficient, specified by r, ranges from −1.0 to +1.0.



Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Spearman's Rank Correlation

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables.

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2 - 1)}$$

$\boldsymbol{\rho}$ = Spearman's rank correlation coefficient
di = Difference between the two ranks of each observation
n = Number of observations

The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

Example

Example : Pearson's correlation Coefficient

ID	Weight (X)	$(X-\bar{X})$	$(X-\bar{X})^2$	Height (Y)	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	148	–6.10	37.21	64	1.00	1.00	–6.10
2	172	17.90	320.41	63	0.00	0.00	0.00
3	203	48.90	2391.21	67	4.00	16.00	195.60
4	109	–45.10	2034.01	60	–3.00	9.00	135.30
5	110	–44.10	1944.81	63	0.00	0.00	0.00
6	134	–20.10	404.01	62	–1.00	1.00	20.10
7	195	40.90	1672.81	59	–4.00	16.00	–163.60
8	147	–7.10	50.41	62	–1.00	1.00	7.10
9	153	–1.10	1.21	66	3.00	9.00	–3.30
10	170	15.90	252.81	64	1.00	1.00	15.90
Σ	1541		9108.90	630		54.00	201.00

$\bar{X} = 1541/10 = 154.10$ $\quad\quad$ $\bar{Y} = 630/10 = 63.00$

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

$$r = \frac{201}{\sqrt{(9108.90)(54)}} \quad r = \frac{201.00}{701.34} = 0.29$$

Example : Spearman's Rank Correlation

Students	Maths	Rank	Science	Rank	d	d square
A	35	3	24	5	2	4
B	20	5	35	4	1	1
C	49	1	39	3	2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4
						14

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

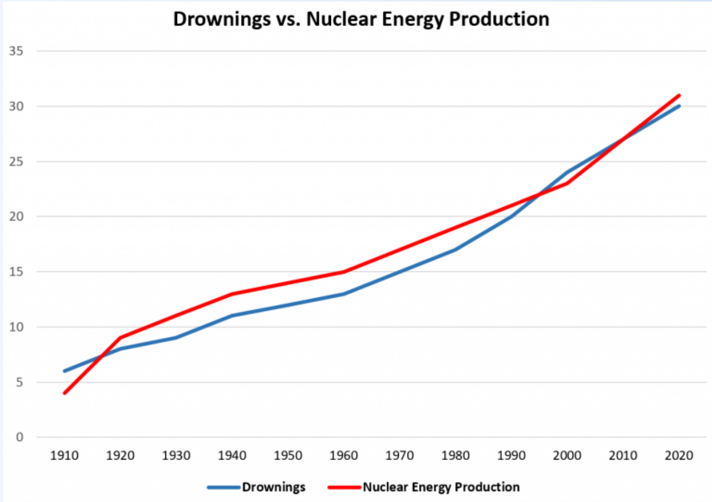
= 1 - (6 * 14) / 5(25 - 1)
= 0.3

Correlation vs Causation

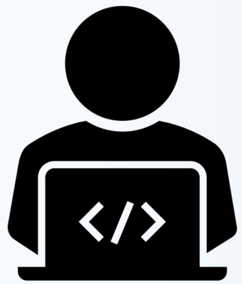
Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.
This is also referred to as cause and effect.

Example :
Pool Drownings vs. Nuclear Energy Production
If we collect data for the total number of pool drownings each year and the total amount of energy produced by nuclear power plants each year, we would find that the two variables are highly correlated.



Note : While causation and correlation can exist at the same time, correlation does not imply causation.



Keep Learning…………… Keep Coding…………… Keep going……………