

STA301 Short Notes Midterm term

Lec 1 to 22

BY vuonlinehelp.blogspot.com

YOUTUBE LINK:

WEBSITE LINK:

[HTTPS://VUONLINEHELP.BLOGSPOT.COM/](https://vuonlinehelp.blogspot.com/)

IMPORTANT NOTE

You should never copy paste the same answer which we are providing. Please do make some changes in the solution file. If you appreciate our work, please share our website vuonlinehelp.blogspot.com with your friends.

آپ کو کبھی بھی وہی جواب کاپی پیسٹ نہیں کرنا چاہیے جو ہم فراہم کر رہے ہیں۔ براہ کرم حل فائل میں کچھ تبدیلیاں کریں۔ اگر آپ ہمارے کام کی تعریف کرتے ہیں تو براہ کرم ہماری ویب سائٹ vuonlinehelp.blogspot.com کو اپنے دوستوں کے ساتھ شیئر کریں

LECTURE NO. 1 Sta301

WHAT IS STATISTICS?

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

Characteristics of Statistics

The important characteristics of Statistics are as follows:

- Statistics are numerically expressed.
- It has an aggregate of facts
- Data are collected in systematic order
- It should be comparable to each other
- Data are collected for a planned purpose

Importance of Statistics

The important functions of statistics are:

- Statistics helps in gathering information about the appropriate quantitative data
- It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
- It provides the exact description and a better understanding
- It helps in designing the effective and proper planning of the statistical inquiry in any field
- It gives valid inferences with the reliability measures about the population parameters from the sample data
- It helps to understand the variability pattern through the quantitative observations

DATA

The word “data” appears in many contexts and frequently is used in ordinary conversation. Although the word carries something of an aura of scientific mystique, its meaning is quite simple and mundane.

Example

In our day to day life, we can collect the following data.

- Number of females per 1000 males in various states of our country.
- Production of wheat in the last 10 years in our country.

- Number of plants in our locality.
- Rainfall in our city in the last 10 years.
- Marks obtained by students.

OBSERVATIONS AND VARIABLES

A **variable** contains all values that measure the same underlying attribute (like height, temperature, duration) across units. An **observation** contains all values measured on the same unit (like a person, or a day, or a city) across attributes.

QUANTITATIVE AND QUALITATIVE VARIABLES

Quantitative data are data about numeric variables (e.g. how many; how much; or how often). **Qualitative data** are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about categorical variables (e.g. what type).

Discrete and continuous variables

A **discrete variable** is a variable that takes on distinct, countable values. A **continuous variable** is a variable that takes on any value within a range, and the number of possible values within that range is infinite.

MEASUREMENT SCALES

long

In Statistics, the variables or numbers are defined and categorized using different scales of measurements.

The four types of scales are:

- Nominal Scale.
- Ordinal Scale.
- Interval Scale.
- Ratio Scale.

Nominal Scale

A **nominal scale** is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

Ordinal Scale

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the “order.” Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

Interval Scale

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

Ratio Scale

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.

ERRORS OF MEASUREMENT

The types ERRORS OF MEASUREMENT are:

- *absolute error*
- *relative error*
- *percentage error*

Absolute Error:

While measuring distances between cities kilometers apart, an error of a few centimeters is negligible and is irrelevant. Consider another case where an error of centimeters when measuring small machine parts is a very significant error. Both the errors are in the order of centimeters but the second error is more severe than the first one.

Relative error

The relative error is defined as the ratio of the absolute error of the measurement to the actual measurement. Using this method we can determine the magnitude of the absolute error in terms of the actual size of the measurement.

Percentage error

Percent error is the difference between estimated value and the actual value in comparison to the actual value and is expressed as a percentage. In other words, the percent error is the relative error multiplied by 100.

LECTURE NO. 2 Sta301

COLLECTION OF DATA

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

PRIMARY AND SECONDARY DATA

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc.

COLLECTION OF PRIMARY DATA

One or more of the following methods are employed to collect primary data:

- Direct Personal Investigation.
- Indirect Investigation.
- Collection through Questionnaires.
- Collection through Enumerators.
- Collection through Local Sources.

DIRECT PERSONAL INVESTIGATION

In direct personal investigation, the investigator obtains the first-hand information from the respondents themselves. He personally visits the respondents to collect the data.

INDIRECT INVESTIGATION

It is a method by which we are gaining the information indirectly by talking to other person who is regarding to have that information.

COLLECTION THROUGH QUESTIONNAIRES

A questionnaire is a list of questions or items used to gather data from respondents about their attitudes, experiences, or opinions. Questionnaires can be used to collect quantitative and/or qualitative information. Questionnaires are commonly used in market research as well as in the social and health sciences.

Collection through Enumerators

In this method, trained enumerators collected the information. They assist the informants in making the entries in the schedules or questionnaires correctly. If the enumerator is well trained, experienced, and discreet, then you can get the most reliable information through this method.

COLLECTION OF SECONDARY DATA

Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc. Always specific to the researcher's needs.

'POPULATION'

A statistical population is the collection of every member of a group possessing the same basic and defined characteristic, but varying in amount or quality from one member to another.

Frame

The frame refers to the list of units (eg, persons, households, businesses, etc) in the survey population. Since the selection of the sample is directly based on this list, the frame is one of the most important tools in the design of a survey.

SAMPLING FRAME

The sampling frame is the list from which units are drawn for the sample. The 'list' may be an actual listing of units, as in a phone book from which phone numbers will be sampled, or some other description of the population, such as a map from which areas will be sampled.

ADVANTAGES OF SAMPLING

Advantages of Sampling	Disadvantages of Sampling
Sampling avoid repetition of query for each and every individual	Selection of good samples is difficult
Sampling gives nearest accurate results	Limited knowledge may mislead the result

SAMPLING & NON-SAMPLING ERRORS

Non-sampling error refers to an error that arises from the result of data collection, which causes the data to differ from the true values. It is different from **sampling error**, which is any difference between the sample values and the universal values that may result from a limited sampling size.

Quota sampling

Quota sampling is a type of non-probability sampling technique. Non-probability sampling focuses on sampling techniques that are based on the judgment of the researcher.

ADVANTAGES OF QUOTA SAMPLING

- There is no need to construct a frame.
- It is a very quick form of investigation.
- Cost reduction.

DISADVANTAGES

- It is a subjective method. One has to choose between objectivity and convenience.
- If random sampling is not employed, it is no longer theoretically possible to evaluate the sampling error.
- Even if the above is not the case, the interviewer may still be making unsuitable selection of sample units.

RANDOM SAMPLING

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population.

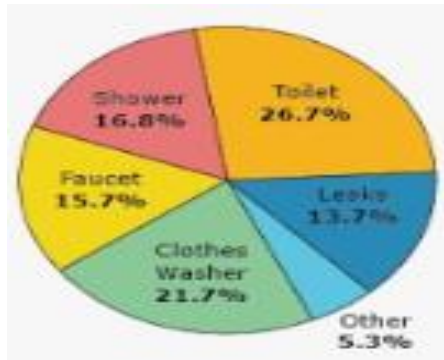
TYPES OF RANDOM SAMPLING

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Cluster Sampling
- Multi-stage Sampling, etc.
-

LECTURE NO. 3 Sta301

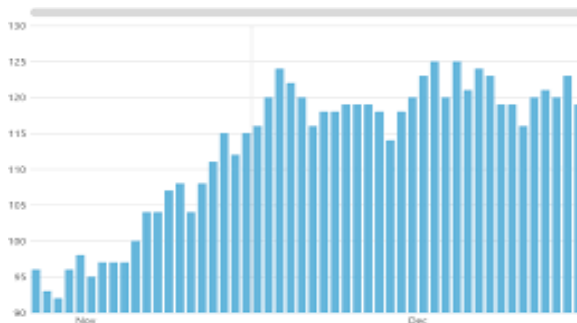
PIE CHART

Pie charts are often used to represent sample data—with data points belonging to a combination of different categories. Each of these categories is represented as a “slice of the pie.”



SIMPLE BAR CHART:

A **simple bar chart** is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis.



BIVARIATE FREQUENCY TABLE

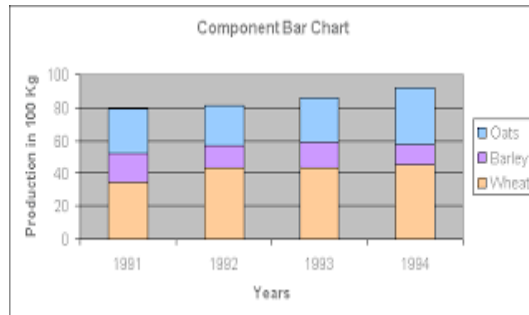
A **table that illustrates the relationship between** two variables by displaying the distribution of one variable across the categories of a second variable.

Table 3.17
Bivariate Frequency Distribution of Age and Marks

Marks	Age in Years				Total
	16 – 18	18 – 20	20 – 22	22 – 24	
10 – 20	2	1	1	-	4
20 – 30	3	2	3	1	9
30 – 40	3	3	5	6	17
40 – 50	2	2	3	4	11
50 – 60	-	1	2	2	5
60 – 70	-	1	2	1	4
Total	10	10	16	14	50

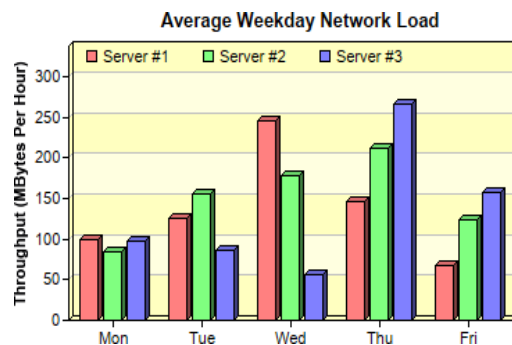
COMPONENT BAR CHART:

A sub-divided or component bar chart is used to represent data in which the total magnitude is divided into different or components.



MULTIPLE BAR CHARTS

A multiple bar graph shows the relationship between different values of data. Each data value is represented by a column in the graph. In a multiple bar graph, multiple data points for each category of data are shown with the addition of columns.



LECTURE NO. 4 Sta301

CONSTRUCTION OF A FREQUENCY DISTRIBUTION

The statistical data that we collect can be presented in the form of a frequency distribution. A frequency distribution refers to summarizing a large data set into a small number of intervals.

Step 1: Sort the data in ascending order

Step 2: Calculate the range of data

Step 3: Decide on the number of intervals in the frequency distribution

Step 4: Determine the intervals.

Step 5: Tally and count the observations under each interval.

$$\text{Relative Frequency} = \frac{\text{Absolute Frequency of Each Interval}}{\text{Total Number of Observations}}$$

HISTOGRAM:

A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

FREQUENCY POLYGON:

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

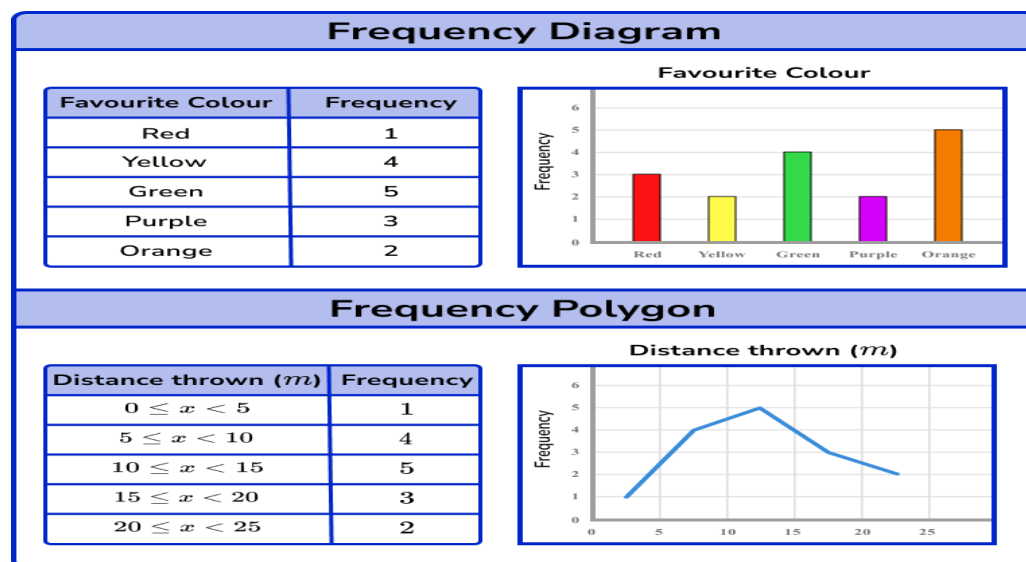
FREQUENCY CURVE:

A Frequency Curve is a smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.

LECTURE NO. 5 Sta301

FREQUENCY POLYGON:

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.



FREQUENCY CURVE

A Frequency Curve is a smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.

VARIOUS TYPES OF FREQUENCY CURVES

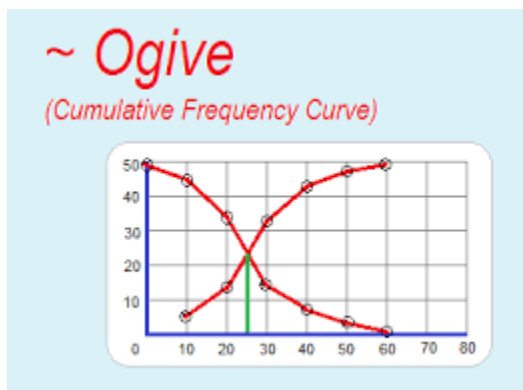
- the symmetrical frequency curve
- the moderately skewed frequency curve
- the extremely skewed frequency curve
- the U-shaped frequency curve

CUMULATIVE FREQUENCY DISTRIBUTION

The cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total.

CUMULATIVE FREQUENCY POLYGON or OGIVE

The cumulative frequency polygon is essentially a line graph drawn on graph paper by plotting actual lower or upper limits of the class intervals on the x-axis and the respective cumulative frequencies of these class intervals on the y-axis.

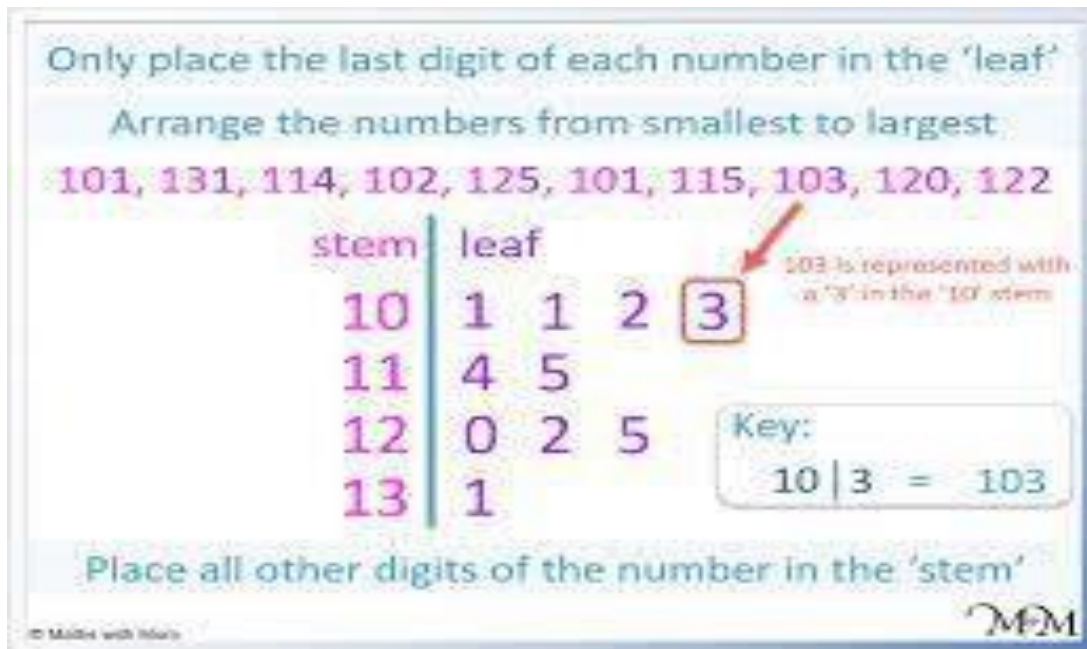


LECTURE NO. 6 Sta301

Stem-and-Leaf Display

A stem-and-leaf display or stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of

a distribution. A stem-and-leaf plot is also called a stem plot, but the latter term often refers to another chart type. A simple stem plot may refer to plotting a matrix of y values onto a common x axis, and identifying the common x value with a vertical line, and the individual y values with symbols on the line.



FREQUENCY DISTRIBUTION polygon graph

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

Class Limits	Class Boundaries	Tally Marks	Frequency
10 – 19	9.5 – 19.5	//	2
20 – 29	19.5 – 29.5	///	3
30 – 39	29.5 – 39.5	###	5
40 – 49	39.5 – 49.5	### /	6
50 – 59	49.5 – 59.5	### /	6
60 – 69	59.5 – 69.5	### /	6
70 - 79	69.5 – 79.5	//	2

EXAMPLE

Construct a stem-and-leaf display for the data of mean annual death rates per thousand at ages 20-65 given below: 7.5, 8.2, 7.2, 8.9, 7.8, 5.4, 9.4, 9.9, 10.9, 10.8, 7.4, 9.7, 11.6, 12.6, 5.0, 10.2, 9.2, 12.0, 9.9, 7.3, 7.3, 8.4, 10.3, 10.1, 10.0, 11.1, 6.5, 12.5, 7.8, 6.5, 8.7,

9.3, 12.4, 10.6, 9.1, 9.7, 9.3, 6.2, 10.3, 6.6, 7.4, 8.6, 7.7, 9.4, 7.7, 12.8, 8.7, 5.5, 8.6, 9.6, 11.9, 10.4, 7.8, 7.6, 12.1, 4.6, 14.0, 8.1, 11.4, 10.6, 11.6, 10.4, 8.1, 4.6, 6.6, 12.8, 6.8, 7.1, 6.6, 8.8, 8.8, 10.7, 10.8, 6.0, 7.9, 7.3, 9.3, 9.3, 8.9, 10.1, 3.9, 6.0, 6.9, 9.0, 8.8, 9.4, 11.4, 10.9

Using the decimal part in each number as the leaf and the rest of the digits as the stem, we get the ordered stem-and-leaf display shown below:

Stem	Leaf
3	9
4	6 6
5	0 4 5
6	0 0 2 2 5 5 6 6 6 8 9
7	1 3 3 3 4 4 5 6 7 7 8 8 8 9
8	1 1 2 4 6 6 7 7 8 8 8 9 9
9	0 1 2 3 3 3 3 4 4 4 6 7 7 9 9
10	0 1 1 2 3 3 4 4 6 6 7 8 8 9 9
11	1 4 4 6 6 9
12	0 1 4 5 6 8 8
14	0

DESCRIPTION OF VARIABLE DATA:

A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.

VARIOUS TYPES OF AVERAGES:

There are three different types of average. These are called the mean, the median, and the mode. Each one gives slightly different information.

The Arithmetic, Geometric and Harmonic means are averages that are mathematical in character, and give an indication of the magnitude of the observed values.

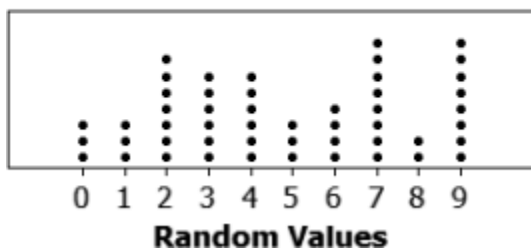
The Median indicates the middle position while the mode provides information about the most frequent value in the distribution or the set of data.

THE MODE: The Mode is defined as that value which occurs most frequently in a set of data i.e. it indicates the most common result.

Dot Plot

a **Dot Plot** is a graph for displaying the distribution of quantitative variable where each dot represents a value. For whole numbers, if a value occurs more than once, the dots are placed one above the other so that the height of the column of dots represents the frequency for that value.

Dotplot of Random Values



Mode

The **mode** is the **value that appears** most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all. Other popular measures of central tendency include the mean, or the average of a set, and the median, the middle value in a set.

$$\hat{X} = l + \left(\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) \times h$$

Where

l = lower class boundary of the modal class,

f_m = frequency of the modal class,

f_1 = frequency of the class preceding the modal class,

f_2 = frequency of the class following modal class, and

h = length of class interval of the modal class

LECTURE NO. 7 Sta301

THE ARITHMETIC MEAN

The arithmetic mean is a measure of central tendency. It allows us to characterize the center of the frequency distribution of a quantitative variable by considering all of the observations with the same weight afforded to each.

$$\bar{X} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

For simplicity, the above formula can be written as $\bar{X} = \frac{\sum X}{n}$

Example:

Class (Mileage Rating)	Frequency (No. of Cars)
30.0 – 32.9	2
33.0 – 35.9	4
36.0 – 38.9	14
39.0 – 41.9	8
42.0 – 44.9	2
Total	30

Class (Mileage Rating)	Class-mark (Midpoint) X
30.0 – 32.9	31.45
33.0 – 35.9	34.45
36.0 – 38.9	37.45
39.0 – 41.9	40.45
42.0 – 44.9	43.45

Class-mark (Midpoint) X	Frequency f	fX
31.45	2	62.9
34.45	4	137.8
37.45	14	524.3
40.45	8	323.6
43.45	2	86.9
	30	1135.5

Applying the formula

$$\bar{X} = \frac{\sum fX}{\sum f},$$

We obtain

$$\bar{X} = \frac{1135.5}{30} = 37.85$$

MEDIAN

The median is the middle number in an ordered data set. The mean is the sum of all values divided by the total number of values.

MEDIAN

The MIDDLE number in a data set

2 4 5 7 12 15 18	3 4 6 10 13 19
↑ Median	↖ $\frac{6+10}{2} = \frac{16}{2} = 8$

52 52 65 73 81 86 89 91 275
Mean = 96 Mode = 52 Median = 81

LECTURE NO. 8 Sta301

Median in Case of a Frequency Distribution of a Continuous Variable:

$$X = l + \frac{h}{f} \left(\frac{n}{2} - c \right)$$

More Files Click and VU WhatsApp Group: [Join Now](#)

Where

- l = lower class boundary of the median class (i.e. that class for which the cumulative frequency is just in excess of $n/2$).
- h = class interval size of the median class
- f = frequency of the median class
- $n = \sum f$ (the total number of observations)
- c = cumulative frequency of the class preceding the median class

EXAMPLE:

Mileage Rating	No. of Cars	Class Boundaries	Cumulative Frequency
30.0 – 32.9	2	29.95 – 32.95	2
33.0 – 35.9	4	32.95 – 35.95	6
36.0 – 38.9	14	35.95 – 38.95	20
39.0 – 41.9	8	38.95 – 41.95	28
42.0 – 44.9	2	41.95 – 44.95	30

$$\begin{aligned}\tilde{X} &= 35.95 + \frac{3}{14}(15 - 6) \\ &= 35.95 + 1.93 \\ &= 37.88 \\ &\approx 37.9\end{aligned}$$

QUARTILES

The quartiles, together with the median, achieve the division of the total area into four equal parts. The first, second and third quartiles are given by the formulae:

1. FIRST QUARTILE

$$Q_1 = l + \frac{h}{f} \left(\frac{n}{4} - c \right)$$

2. second QUARTILE (I.E. MEDIAN)

$$Q_2 = l + \frac{h}{f} \left(\frac{2n}{4} - c \right) = l + \frac{h}{f} (n/2 - c)$$

3. THIRD QUARTILE

$$Q_3 = l + \frac{h}{f} \left(\frac{3n}{4} - c \right)$$

DECILES & PERCENTILES

The deciles and the percentiles give the division of the total area into 10 and 100 equal parts respectively. The formula for the first decile is:

$$D_1 = l + \frac{h}{f} \left(\frac{n}{10} - c \right)$$

$$D_2 = l + \frac{h}{f} \left(\frac{2n}{10} - c \right)$$

$$D_3 = l + \frac{h}{f} \left(\frac{3n}{10} - c \right)$$

It is easily seen that the 5th decile is the same quantity as the median. The formula for the first percentile is:

$$P_1 = l + \frac{h}{f} \left(\frac{n}{100} - c \right)$$

$$P_2 = l + \frac{h}{f} \left(\frac{2n}{100} - c \right)$$

$$P_3 = l + \frac{h}{f} \left(\frac{3n}{100} - c \right)$$

LECTURE NO. 9 Sta301

GEOMETRIC MEAN

In Mathematics, the Geometric Mean (GM) is the average value or mean which signifies the central tendency of the set of numbers by finding the product of their values.

Basically, we multiply the numbers altogether and take the nth root of the multiplied numbers, where n is the total number of data values.

$$G = \sqrt[n]{X_1 X_2 \dots X_n}$$

$$\begin{aligned}\log G &= \frac{1}{n} [\log X_1 + \log X_2 + \dots + \log X_n] \\ &= \frac{\sum \log X}{n} \\ G &= \text{antilog} \left[\frac{\sum \log X}{n} \right]\end{aligned}$$

EXAMPLE:

Find the geometric mean of numbers:

45, 32, 37, 46, 39, 36, 41, 48, 36

Solution:

We need to compute the numerical value of

$$= \sqrt[9]{45 \times 32 \times 37 \times 46 \times 39 \times 36 \times 41 \times 48 \times 36}$$

But, obviously, it is a bit cumbersome to find the ninth root of a quantity. So we make use of logarithms, as shown below:

X	log X
45	1.6532
32	1.5052
37	1.5682
46	1.6628
39	1.5911
36	1.5563
41	1.6128
48	1.6812
36	1.5563
	14.3870

$$\begin{aligned}\log G &= \frac{\sum \log X}{n} \\ &= \frac{14.3870}{9} = 1.5986\end{aligned}$$

$$\begin{aligned}\text{Hence } G &= \text{anti log } 1.5986 \\ &= 39.68\end{aligned}$$

GEOMETRIC MEAN FOR GROUPED DATA

In case of a frequency distribution having k classes with midpoints $X_1, X_2, \dots, X_k$ and the corresponding frequencies $f_1, f_2, \dots, f_k$ (such that $\sum f_i = n$), the geometric mean is given by

$$G = \sqrt[n]{X_1^{f_1} X_2^{f_2} \dots X_k^{f_k}}$$
$$\log G = \frac{1}{n} \left[f_1 \log X_1 + f_2 \log X_2 + \dots + f_k \log X_k \right]$$
$$= \frac{\sum f \log X}{n}$$
$$G = \text{anti log} \left[\frac{\sum f \log X}{n} \right]$$

RELATION BETWEEN ARITHMETIC, GEOMETRIC AND HARMONIC MEANS

MID-RANGE

If there are n observations with x_0 and x_m as their smallest and largest observations respectively, then their mid-range is defined as

$$\text{mid - range} = \frac{x_0 + x_m}{2}$$

MID-QUARTILE RANGE

If $x_1, x_2, \dots, x_n$ are n observations with Q_1 and Q_3 as their first and third quartiles respectively, then their mid-quartile range is defined as

$$\text{mid - quartile range} = \frac{Q_1 + Q_3}{2}$$

ABSOLUTE VERSUS RELATIVE

There are two types of measurements of dispersion: absolute and relative. An absolute measure of dispersion is one that measures the dispersion in terms of the same units or in the square of units, as the units of the data. For example, if the units of the data are rupees, meters, kilograms, etc., the units of the measures of dispersion will also be rupees, meters, kilograms, etc. On the other hand, relative measure of dispersion is one that is expressed in the form of a ratio, co-efficient of percentage and is independent of the units of measurement.

RANGE

The range is defined as the difference between the two extreme values of a data-set, i.e. $R = X_m - X_0$ where X_m represents the highest value and X_0 the lowest. Evidently, the calculation of the range is a simple question of MENTAL arithmetic.

Disadvantages

1. It *ignores* all the INFORMATION available from the intermediate observations.
2. It might give a MISLEADING picture of the spread in the data.

From THIS point of view, it is an unsatisfactory measure of dispersion. However, it is APPROPRIATELY used in statistical quality control charts of manufactured products, daily temperatures, stock prices, etc.

$$\begin{aligned} & \frac{(Midrange - X_0) + (X_m - Midrange)}{2} \times 2 \\ &= \frac{Midrange - X_0 + X_m - Midrange}{2} \\ &= \frac{X_m - X_0}{2} \end{aligned}$$

The range is an *absolute* measure of dispersion. Its *relative* measure is known as the *Coefficient Of Dispersion*, and is defined by the relation given below:

$$= \frac{\frac{1}{2} (Range)}{Mid - Range}$$

$$= \frac{\frac{X_m - X_0}{2}}{\frac{X_m + X_0}{2}} = \frac{X_m - X_0}{X_m + X_0}$$

$$\overline{X_m + X_0}$$

QUARTILE DEVIATION

The quartile deviation is defined as half of the difference between the third and first quartiles i.e.

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

COEFFICIENT OF QUARTILE DEVIATION

$$\begin{aligned} &= \frac{\text{Quartile Deviation}}{\text{Mid - Quartile Range}} \\ &= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}, \end{aligned}$$

MEAN DEVIATION

Mean deviation is a measure of dispersion, which is known as the average deviation. Mean deviation can be computed from the mean or median. Mean deviation is the arithmetic deviation of different items of central tendency. It may be the mean or the median.

$$M.D. = \frac{\sum |d|}{N}$$

LECTURE NO. 11 Sta301

MEAN DEVIATION

The **mean deviation** is the mean of the absolute deviations of the observations or values from a suitable average. This suitable average may be the mean, median or mode. We also know it as the mean absolute deviation.

The basic formula to calculate mean deviation for a given data set is as follows:

$$\text{Mean Deviation} = \frac{\sum |X - \bar{X}|}{N}$$

Where,

X = denotes each value in the data set
 $\overline{X}$ = denotes the mean value of the data set
 N = total number of data values
 $||$ = represents absolute value, i.e. it ignores the sign

MEAN DEVIATION FOR GROUPED DATA

Ungrouped data is data that has not been sorted or categorized into groups and is still in its raw form. Generally, ungrouped data includes individual data series. The formula to calculate mean deviation for ungrouped data is as follows:

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - \bar{x}|}{n}$$

Where,

x_i = i^{th} observation

$\overline{X}$ = central point of the data (mean, median or mode)

n = number of observations

Variance

The sum of squares of the deviations of the X from the mean, divided by the no. of values.

$$\text{STANDARD DEVIATION} = \frac{\sum (x - \bar{x})^2}{n}$$

Variance is equal to the average squared deviations from the mean, while standard deviation is the number's square root. Also, the standard deviation is a square root of variance.

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

STANDARD DEVIATION IN CASE OF GROUPED DATA

$$S = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$$

SHORT CUT FORMULA OF THE STANDARD DEVIATION IN CASE OF GROUPED DATA

$$S = \sqrt{\left\{ \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n} \right)^2 \right\}}$$

COEFFICIENT OF VARIATION

The coefficient of variation (CV) is the ratio of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population. The higher the CV, the greater the dispersion.

$$C.V. = \frac{S}{\bar{X}} \times 100$$

LECTURE NO. 12 Sta301

CHEBYCHEV'S THEOREM

Chebyshev's Theorem estimates the minimum proportion of observations that fall within a specified number of standard deviations from the mean. This theorem applies to a broad range of probability distributions. Chebyshev's Theorem is also known as Chebyshev's Inequality.

Here are two forms of the equation. One determines how close to the mean the data lie and the other calculate how far away from the mean they fall:

Maximum proportion of observations that are more than k standard deviations from the mean

$$\frac{1}{k^2}$$

Minimum proportion of observations that are within k standard
deviations of the mean

$$1 - \frac{1}{k^2}$$

Where k equals the number of standard deviations in which you are interested. K must be greater than 1.

Five-number summaries

A **five-number summary** is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q_1), median value (Q_2), upper quartile (Q_3), maximum value.

LECTURE NO. 13 Sta301

BOX AND WHISKER PLOT

A **box and whisker plot** is defined as a graphical method of displaying variation in a set of data. In most cases, a histogram analysis provides a sufficient display, but a box and whisker plot can provide additional detail while allowing multiple sets of data to be displayed in the same graph.

The procedure to develop a box and whisker plot comes from the five statistics below. You can also download the box and whisker plot template.

1. **Minimum value:** The smallest value in the data set
2. **Second quartile:** The value below which the lower 25% of the data are contained
3. **Median value:** The middle number in a range of numbers
4. **Third quartile:** The value above which the upper 25% of the data are contained
5. **Maximum value:** The largest value in the data set

INTERPRETATION OF THE BOX AND WHISKER PLOT

In a box and whisker plot: The left and right sides of the box are the lower and upper quartiles. The box covers the interquartile interval, where 50% of the data is found. The vertical line that split the box in two is the median.

COEFFICIENT OF SKEWNESS

The coefficient of skewness measures the skewness of a distribution. It is based on the notion of the moment of the distribution. This coefficient is one of the measures of skewness.

$$\text{skewness} = \frac{\bar{x} - \text{mode}}{S},$$

LECTURE NO. 14 Sta301

PEARSON'S COEFFICIENT OF SKEWNESS:

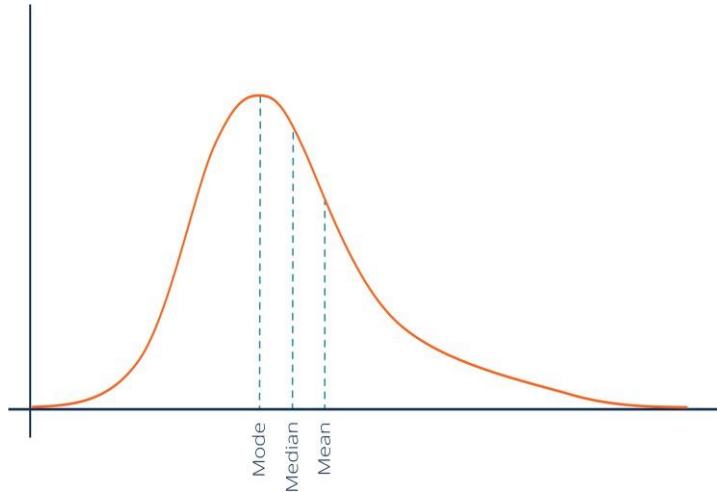
pearson's coefficient of skewness (second method) is calculated by multiplying the difference between the mean and median, multiplied by three. The result is divided by the standard deviation. You can use the Excel functions AVERAGE, MEDIAN and STDEV.P to get a value for this measure.

In Power BI, this is implemented by the following formula:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

POSITIVE SKEWNESS

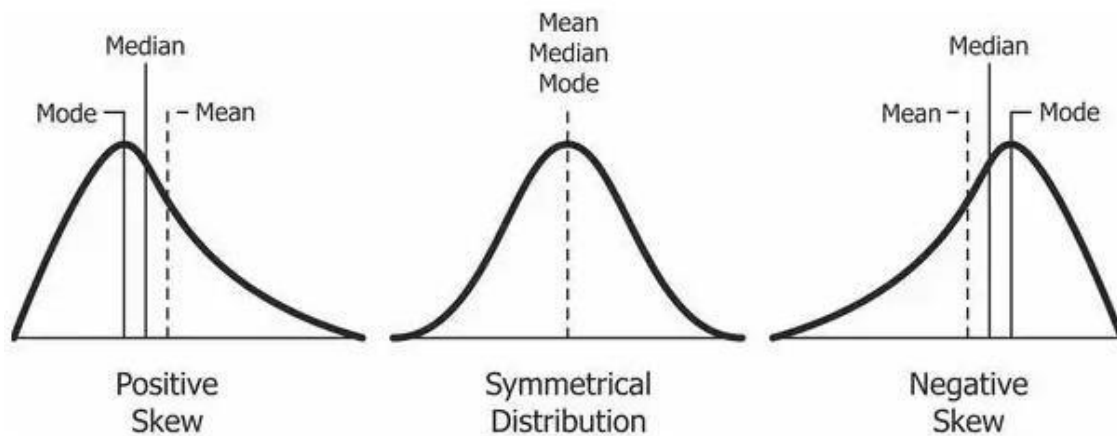
In statistics, a positively skewed (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer. The positively skewed distribution is the direct opposite of the negatively skewed distribution.



Mean > Median > Mode

NEGATIVE SKEWNESS

Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.

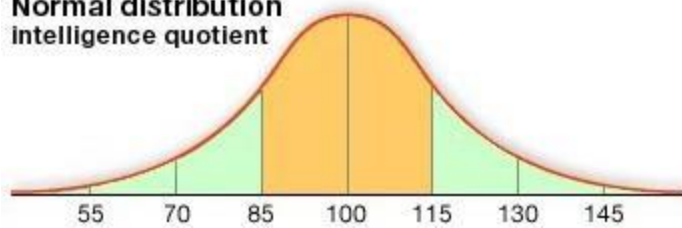


$3 * (\text{Mean} - \text{Median}) / \text{Standard Deviation}$.

Relative

A relative frequency distribution shows the proportion of the total number of observations associated with each value or class of values and is related to a probability distribution, which is extensively used in statistics.

Normal distribution
intelligence quotient



Hint: To find the Bowley's coefficient of skewness, we first need to calculate three quartiles at $\frac{N}{4}$,

$\frac{N}{2}$ and $\frac{3N}{4}$ respectively. Then, we can calculate the Bowley's coefficient of skewness using the

formula
$$SK_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

KURTOSIS

Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur. **Excess kurtosis** is the tailedness of a distribution relative to a normal distribution.

- Distributions with medium kurtosis (medium tails) are **mesokurtic**.
- Distributions with low kurtosis (thin tails) are **platykurtic**.
- Distributions with high kurtosis (fat tails) are **leptokurtic**.

Tails are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, tails represent how often **outliers** occur.

Leptokurtic

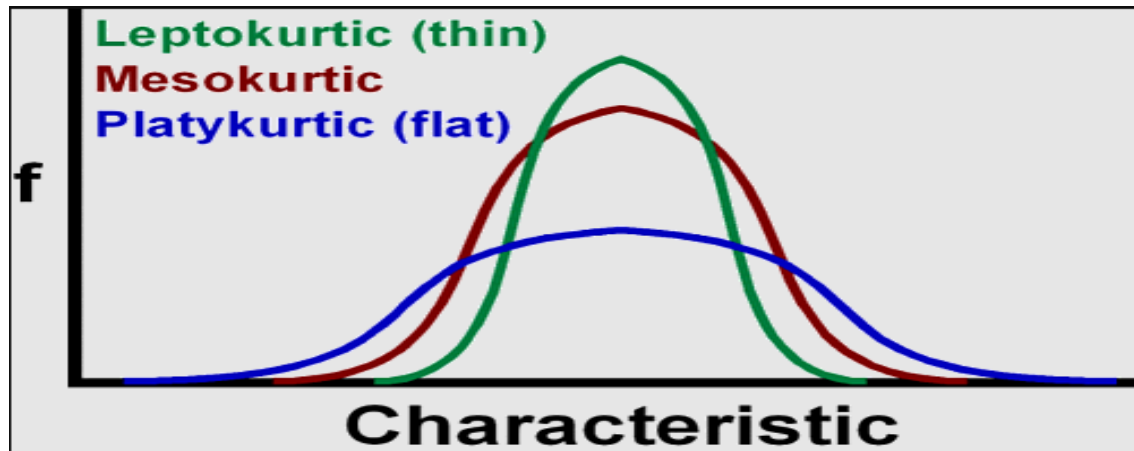
Leptokurtic distributions are statistical distributions with kurtosis greater than three. It can be described as having a wider or flatter shape with fatter tails resulting in a greater chance of extreme positive or negative events. It is one of three major categories found in kurtosis analysis.

Platykurtic

The term "platykurtic" refers to a statistical distribution in which the excess kurtosis value is **negative**. For this reason, a platykurtic distribution will have thinner tails than a normal distribution will, resulting in fewer extreme positive or negative events.

Mesokurtic

Mesokurtic is a statistical term used to describe the outlier characteristic of a probability distribution that is close to zero. Mesokurtic distributions are similar to normal distributions, in which extreme or outlier events are very unlikely.



MOMENTS

The third moment is called skewness, and the fourth moment is known as kurtosis. The third moment measures the asymmetry of distribution while the fourth moment measures how heavy the tail values are. Physicists generally use the higher-order moments in applications of physics.

$$\frac{1}{n} \sum (x_i - \bar{x})^l = \frac{1}{n} \sum (x_i - \bar{x})$$

Is called the first sample moment about the mean, and is denoted by m_1 . Similarly, the quantity.

$$\frac{1}{n} \sum (x_i - \bar{x})^2$$

Is called the second sample moment about the mean, and is denoted by m_2 .

$$m_r = \frac{1}{n} \sum (x_i - \bar{x})^r$$

Moments about the mean are also called the central moments or the mean moments.

$$m'_r = \frac{1}{n} \sum (x_i - \alpha)^r \quad \text{for sample data}$$

For $r = 1$, we have

$$m_1 = \frac{1}{n} \sum (x_i - \bar{x}) = \frac{\sum x_i}{n} - \bar{x} = \bar{x} - \bar{x} = 0,$$

and

$$m'_1 = \frac{1}{n} \sum (x_i - \alpha) = \frac{\sum x_i}{n} - \alpha = \bar{x} - \alpha.$$

Putting $r = 2$ in the relation for mean moments, we see that

$$m_2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$m'_r = \frac{1}{n} \sum (x_i - \alpha)^r$$

if we put $\alpha = 0$, we obtain

$$m'_r = \frac{1}{n} \sum x_i^r$$

SHEPPARD'S CORRECTIONS:

In statistics, **Sheppard's corrections** are approximate corrections to estimates of moments computed from binned data. The concept is named after William Fleetwood Sheppard.

Let m_k be the measured k^{th} moment, $\hat{\mu}_k$ the corresponding corrected moment, and c the breadth of the class interval. No correction is necessary for the mean (first moment about zero). The first few measured and corrected moments about the mean are then related as follows:

$$\hat{\mu}_2 = m_2 - \frac{1}{12}c^2$$

$$\hat{\mu}_3 = m_3$$

$$\hat{\mu}_4 = m_4 - \frac{1}{2}m_2c^2 + \frac{7}{240}c^4.$$

The reason for doing so is that, in many situations, it is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the mean-moments using the relationships that I am now going to convey to you. The equations are:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2;$$

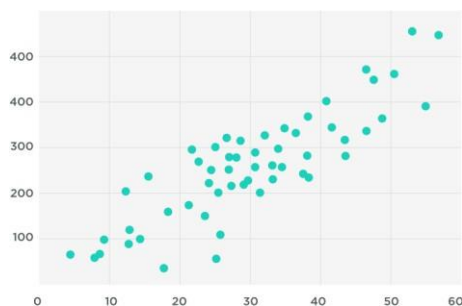
$$m_3 = m'_3 - 3 m'_2 m'_1 + 2(m'_1)^3, \text{ and}$$

$$m_4 = m'_4 - 4 m'_3 m'_1 + 6 m'_2 (m'_1)^2 - 3(m'_1)^4$$

LECTURE NO. 15 Sta301

Scatter plot

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.



Scatter formula

Step 1: Take any two points on the line; preferably the points that are at the intersection of the grid. We can take any points on the line but it gets easier with the calculations if

we take a point at the intersection of the grid. Find the coordinates of the points we identified.

Step 2: Find the slope of the line using the points.

Find the slope of the line by substituting the value of the coordinates in the equation

$\text{slope}(m) = \frac{y_2 - y_1}{x_2 - x_1}$, where (x_1, y_1) are the coordinates of one point and (x_2, y_2) are the coordinates of second point.

Step 3: Find the y-intercept. To find the y- intercept (b), take any point identified in step 1 and substitute the value of the x-coordinate and the y-coordinate of the point and the value of slope (m) found in step 2 in the equation $y = mx + b$.

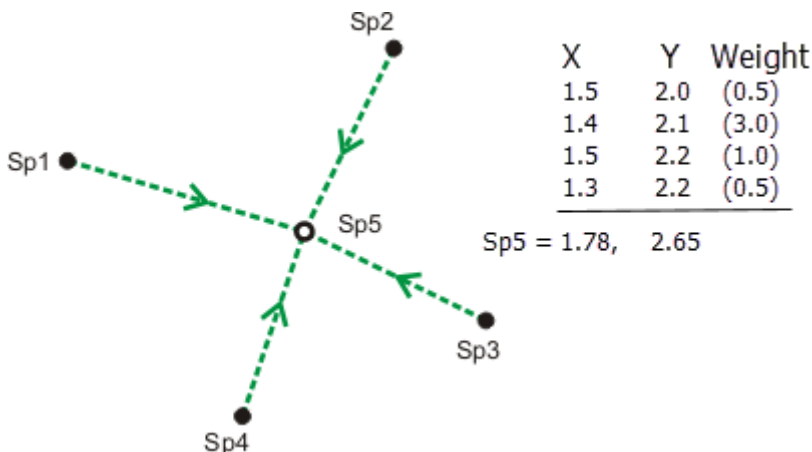
Step 4: Find the equation of a given line of best fit for the given scatter plot.

The equation of a line of fit is $y = mx + b$, where m is the slope found in step 2 and b is the y-intercept found in step 3.

Step 5: To make predictions, take a look at the given word problem and substitute the given value in the equation of the line of fit we found in the previous step to predict the unknown value asked.

THE PRINCIPLE OF LEAST SQUARES

A least-squares adjustment uses statistical analysis to estimate the most likely coordinates for connected points in a measurement in a network. In ArcGIS Pro, a least-squares adjustment can be run on a parcel fabric using the following tools: Analyze Parcels By Least Squares Adjustment tool.



$$\left. \begin{aligned} \sum Y &= na + b \sum X \\ \sum XY &= a \sum X + b \sum X^2 \end{aligned} \right\}$$

STANDARD ERROR OF ESTIMATE

The **standard error of the estimate** is a measure of the average deviation of the errors, the difference between the $\hat{y}$ -values predicted by the multiple regression model and the y - values in the sample. The standard error of the estimate for the regression model is the standard deviation of the errors/residuals.

$$s_{y.x} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}}$$
$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n - 2}}$$

Step 1: Note the number of measurements (n) and determine the sample mean (μ). It is the average of all the measurements.

Step 2: Determine how much each measurement varies from the mean.

Step 3: Square all the deviations determined in step 2 and add altogether: $\sum (x_i - \mu)^2$

Step 4: Divide the sum from step 3 by one less than the total number of measurements ($n-1$).

Step 5: Take the square root of the obtained number, which is the standard deviation (σ).

Step 6: Finally, divide the standard deviation obtained by the square root of the number of measurements (n) to get the standard error of your estimate.

Go through the example given below to understand the method of calculating standard error.

INTERPRETATION OF $S_{y.x}$

$S_{y.x}$ is computed in a very similar way, but the denominator is $n-K$, where K is the number of parameters fit by regression. The value $n-K$ is the number of degrees of freedom of the regression. If you only fit one parameter, then the RMSE and $S_{y.x}$ are the same.

CORRELATION

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

PEARSON'S COEFFICIENT OF CORRELATION

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best

method of measuring the association between variables of interest because it is based on the method of covariance.

where, covariance of X and Y is defined as
$$r = \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}}$$

$$Cov(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$

SHORT CUT FORMULA FOR THE PEARSON'S COEFFICIENT OF CORRELATION

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

LECTURE NO. 16 Sta301

“SET”

A set is the mathematical model for a collection of different things a set contains elements or members, which can be mathematical objects of any kind: numbers, symbols, points in space, lines, other geometrical shapes, variables, or even other sets. The set with no element is the empty set; a set with a single element is a singleton. A set may have a finite number of elements or be an infinite set. Two sets are equal if they have precisely the same elements.

A = {a, b, c, d} or
B = {1, 2, 3, 7}

Type of sets

Empty Sets

The set, which has no elements, is also called a null set or void set. It is denoted by {}. Below are the two examples of the empty set.

Singleton Sets

The set which has just one element is named a singleton set.
For example, Set A = { 8 } is a singleton set.

Finite and Infinite Sets

A set that has a finite number of elements is known as a finite set, whereas the set whose elements can't be estimated, but has some figure or number, which is large to precise in a set, is known as infinite set.

For example, set $A = \{3,4,5,6,7\}$ is a finite set, as it has a finite number of elements.

Equal Sets

If every element of set A is also the elements of set B and if every element of set B is also the elements of set A, then sets A and B are called equal sets. It means set A and set B have equivalent elements and that we can denote it as:

$$A = B$$

For example, let $A = \{3,4,5,6\}$ and $B = \{6,5,4,3\}$, then $A = B$

Subsets

A set S is said to be a subset of set T if the elements of set S belong to set T, or you can say each element of set S is present in set T. Subset of a set is denoted by the symbol ($\subset$) and written as $S \subset T$.

We can also write the subset notation as:

$$S \subset T \text{ if } p \in S \Rightarrow p \in T$$

Power Sets

The set of all subsets is known as power sets. We know the empty set is a subset of all sets, and each set is a subset of itself. Taking an example of set $X = \{2,3\}$. From the above-given statements, we can write,

$\{\}$ is a subset of $\{2,3\}$

$\{2\}$ is a subset of $\{2,3\}$

$\{3\}$ is a subset of $\{2,3\}$

$\{2,3\}$ is also a subset of $\{2,3\}$

Therefore, power set of $X = \{2,3\}$,

$$P(X) = \{\{\}, \{2\}, \{3\}, \{2,3\}\}$$

Universal Sets

A set that contains all the elements of other sets is called a universal set. Generally, it is represented as 'U.'

For example, set $A = \{1,2,3\}$, set $B = \{3,4,5,6\}$, and $C = \{5,6,7,8,9\}$.

Disjoint Sets

If two sets X and Y do not have any common elements, and their intersection results in zero(0), then set X and Y are called disjoint sets. It can be represented as, $X \cap Y = 0$.

OPERATIONS ON SETS

Union of Sets

The union of two sets consists of all their elements. It is denoted by (U).

For example, set $A = \{2,3,7\}$ and set $B = \{4,5,8\}$.

Then the union of set A and set B will be:

$$A \cup B = \{2,3,7,4,5,8\}$$

Intersection of Sets

The set of all elements, which are common to all the given sets, gives an intersection of sets. It is denoted by $\cap$.

For example, set $A = \{2,3,7\}$ and set $B = \{2,4,9\}$.

$$\text{So, } A \cap B = \{2\}$$

Difference of Sets

The difference between set S and set T is such that it has only those elements which are in the set S and not in the set T. $S - T = \{p : p \in S \text{ and } p \notin T\}$

Similarly, $T - S = \{p : p \in T \text{ and } p \notin S\}$.

Complement of a Set

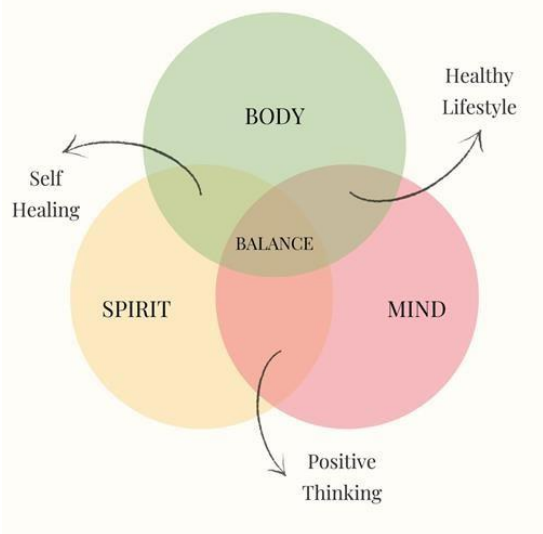
Let U be the universal set and let $A \subset U$. Then, the complement of A, denoted by A' or $(U - A)$, is defined as

$$A' = \{x \in U : x \notin A\}$$

$$X \cap A' = \emptyset$$

VENN DIAGRAM

A Venn diagram is an illustration that uses circles to show the relationships among things or finite groups of things. Circles that overlap have a commonality while circles that do not overlap do not share those traits. Venn diagrams help to visually represent the similarities and differences between two concepts.



COMPLEMENTATION:

The **complement of set A** is defined as a set that contains the elements present in the universal set but not in set A. For example, Set $U = \{2,4,6,8,10,12\}$ and set $A = \{4,6,8\}$, then the complement of set A, $A' = \{2,10,12\}$.

Commutative laws

$$A \cup B = B \cup A \text{ and}$$

$$A \cap B = B \cap A$$

Associative laws

$$(A \cup B) \cup C = A \cup (B \cup C)$$

and

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive laws

$$A \cap (B \cup C)$$

$$= (A \cap B) \cup (A \cap C)$$

$$\text{and } A \cup (B \cap C)$$

$$= (A \cup B) \cap (A \cup C)$$

Idempotent laws

$$A \cup A = A \text{ and}$$

$$A \cap A = A$$

Identity laws

$$A \cup S = S,$$

$$A \cap S = A,$$

$$A \cup \phi = A, \text{ and}$$

$$A \cap \phi = \phi.$$

Complementation laws

$$A \cup \bar{A} = S,$$

$$A \cap \bar{A} = \phi,$$

$$(\bar{\bar{A}}) = A,$$

$$\bar{\bar{S}} = \phi, \text{ and}$$

$$\bar{\phi} = S$$

De Morgan's laws

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B},$$

$$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$$

PARTITION OF SETS

A partition of a set S is a sub-division of the set into non-empty subsets that are disjoint and exhaustive, i.e. their union is the set S itself.

This implies that:

- i) $A_i \cap A_j = \phi$, where $i \neq j$;
- ii) $A_1 \cap A_2 \cup \dots \cup A_n = S$.

The subsets in a partition are called cells.

CLASS OF SETS

A set of sets is called a class. For example, in a set of lines, each line is a set of points.

POWER SET

The class of ALL subsets of a set A is called the Power Set of A and is denoted by P(A).

For example, if $A = \{H, T\}$, then $P(A) = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

CARTESIAN PRODUCT OF SETS

The **Cartesian product of sets** A and B, denoted by $A \times B$, (read as “A cross B”), is a set that contains all ordered pairs (x, y) , where x belongs to A and y belongs to B.

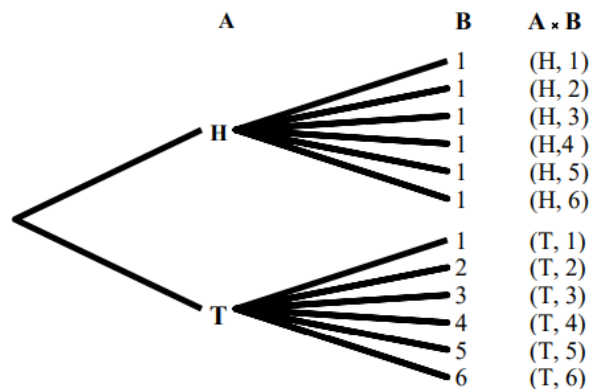
Symbolically, we write

$$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}$$

TREE DIAGRAM

In probability theory, a **tree diagram** may be used to represent a probability space. Tree diagrams may represent a series of independent events (such as a set of coin flips) or conditional probabilities (such as drawing cards from a deck, without replacing the cards). Each node on the diagram represents an event and is associated with the probability of that event. The root node represents the certain event and therefore has probability 1. Each set of sibling nodes represents an exclusive and exhaustive partition of the parent event.

Tree Diagram



LECTURE NO. 17 Sta301

COUNTING RULES

The **Fundamental Counting Principle** (also called the counting rule) is a way to figure out the number of outcomes in a probability problem. Basically, you multiply the events together to get the total number of outcomes.

RULE OF PERMUTATION

The special permutation rule states that anything permute itself is equivalent to itself factorial.

Example: $(3)3 = 3!$

$${}^n P_r = n(n-1)(n-2) \dots (n-r+1)$$

$$= \frac{n!}{(n-r)!}$$

$$\text{is } P = \frac{n!}{n_1! n_2! \dots n_k!}$$

(where $\sum n_i = n$)

RULE OF COMBINATION

Dempster's rule of combination is a rule for combining two or more belief functions; when the belief functions combined are based on distinct or “independent” sources of evidence, the rule corresponds intuitively to the pooling of evidence.

and is given by

$${}^n C_r \text{ or } \binom{n}{r},$$
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

It should be noted that

$${}^n P_r = r! \binom{n}{r}$$

RANDOM EXPERIMENT

The Random Experiment: A Precise Definition of a Random Situation. A random experiment is any well-defined procedure that produces an observable outcome that could not be perfectly predicted in advance. A random experiment must be well defined to eliminate any vagueness or surprise.

PROPERTIES OF A RANDOM EXPERIMENT

A random experiment is a process characterized by the following properties:

- (i) It is performed according to some set of rules,
- (ii) It can be repeated arbitrarily often,

- (iii) The result of each performance depends on chance and cannot be predicted uniquely.

Sample space

A **sample space** is the set of all possible outcomes of a statistical experiment, and it is sometimes referred to as a probability space. And outcomes are observations of the experiment, and they are sometimes referred to as sample points.

Events

In **probability**, **events** are the outcomes of an experiment. The probability of an event is the measure of the chance that the event will occur as a result of an experiment.

SIMPLE & COMPOUND EVENTS

Simple events are events where one experiment happens at a time and it will have a single outcome such as tossing a coin. **Compound events** are events where there is more than one possible outcome such as roll a five using a 6-sided die.

OCCURRENCE OF AN EVENT

An event A is said to occur if and only if the outcome of the experiment corresponds to some element of A.

COMPLEMENTARY EVENT

In probability theory, the complement of any event A is the event [not A], i.e. the event that A does not occur. The event A and its complement [not A] are mutually exclusive and exhaustive.

MUTUALLY EXCLUSIVE EVENTS

In statistics and probability theory, two events are mutually exclusive if they cannot occur at the same time. The simplest example of mutually exclusive events is a coin toss. A tossed coin outcome can be either head or tails, but both outcomes cannot occur simultaneously.

EXHAUSTIVE EVENTS

A set of events are called **exhaustive events** if at least one of them necessarily occurs whenever the experiment is performed. Also, the union of all these events constitutes the sample space of that experiment.

PARTITION OF THE SAMPLE SPACE

A collection of sets $B_1, B_2, \dots, B_n$ is said to partition the sample space if the sets

- (i) are mutually disjoint and
- (ii) Have as union the entire sample space.

A simple example of a partition is given by a set B, together with its complement B^c .

EQUALLY LIKELY EVENTS

Equally likely events are events that have the same theoretical probability (or likelihood) of occurring. Each numeral on a die is equally likely to occur when the die is tossed.

LECTURE NO. 18 Sta301

MUTUALLY EXCLUSIVE EVENTS

In statistics and probability theory, two events are mutually exclusive if they cannot occur at the same time. The simplest example of mutually exclusive events is a coin toss. A tossed coin outcome can be either head or tails, but both outcomes cannot occur simultaneously.

EXHAUSTIVE EVENTS

Set of events are called exhaustive events if at least one of them necessarily occurs whenever the experiment is performed. Also, the union of all these events constitutes the sample space of that experiment.

PARTITION OF THE SAMPLE SPACE

Partitions: A collection of sets $B_1, B_2, \dots, B_n$ is said to partition the sample space if the sets (i) are mutually disjoint and (ii) have as union the entire sample space. A simple example of a partition is given by a set B, together with its complement B^c .

EQUALLY LIKELY EVENTS

Equally likely events are **events that have the same theoretical probability (or likelihood) of occurring**. Each numeral on a die is equally likely to occur when the die is tossed.

SUBJECTIVE OR PERSONALISTIC PROBABILITY

Subjective probability is a type of probability derived from an individual's personal judgment or own experience about whether a specific outcome is likely to occur. It contains no formal calculations and only reflects the subject's opinions and past experience.

THE CLASSICAL OR 'A PRIORI' DEFINITION OF PROBABILITY

A priori probability refers to the likelihood of an event occurring when there is a finite amount of outcomes and each is equally likely to occur. The outcomes in a priori probability are not influenced by the prior outcome.

$$P(A) = \frac{m}{n}$$
$$= \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

LECTURE NO. 19 Sta301

STATISTICAL PROBABILITY

The probability is the measure of the likelihood of an event to happen. It measures the certainty of the event. The formula for probability is given by;

$P(E) = \text{Number of Favourable Outcomes} / \text{Number of total outcomes.}$

$P(E) = n(E)/n(S)$

Here,

$n(E)$ = Number of event favourable to event E

$n(S)$ = Total number of outcomes

LAW OF COMPLEMENTATION

The union of a set A and its complement A' gives the universal set U of which, A and A' are a subset. $A \cup A' = U$. Also, the intersection of a set A and its complement A' gives the empty set $\emptyset$. $A \cap A' = \emptyset$

For Example: If $U = \{1, 2, 3, 4, 5\}$ and $A = \{1, 2, 3\}$ then $A' = \{4, 5\}$.

$$P(\bar{A}) = 1 - P(A).$$

ADDITION LAW

The addition rule states the probability of two events is the sum of the probabilities of two events that will happen minus the probability of both the events that will happen. Mathematically, the addition rule of probability is expressed as:

$P(A \cup B) = P(A) + P(B) - P(A \cap B).$

LECTURE NO. 20 Sta301

ADDITION LAW

The addition rule states the probability of two events is the sum of the probabilities of two events that will happen minus the probability of both the events that will happen. Mathematically, the addition rule of probability is expressed as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

COROLLARY-1

If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

(Since $A \cap B$ is an impossible event, hence $P(A \cap B) = 0$.)

COROLLARY-2

If $A_1, A_2, \dots, A_k$ are k mutually exclusive events, then the probability that one of them occurs, is the sum of the probabilities of the separate events, i.e.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

Conditional probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Where $P(B) > 0$

(If $P(B) = 0$, the conditional probability $P(A/B)$ remains undefined.)

Similarly

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

MULTIPLICATION LAW

(The probability of A given B equals the probability of A and B divided by the probability of B.)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

LECTURE NO. 21 Sta301

INDEPENDENT EVENTS

Independent events are those events whose occurrence is not dependent on any other event. For example, if we flip a coin in the air and get the outcome as Head, then again if we flip the coin but this time we get the outcome as Tail. In both cases, the occurrence of both events is independent of each other.

If A and B are independent events, then $P(A | B) = P(A)$

Using the Multiplication rule of probability, $P(A \cap B) = P(B) \cdot P(A | B)$

$$P(A \cap B) = P(B) \cdot P(A)$$

LECTURE NO. 22 Sta301

BAYES' THEOREM

In probability theory and statistics, **Bayes' theorem** (alternatively **Bayes' law** or **Bayes' rule**), named after Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

For example, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on their age) than simply assuming that the individual is typical of the population as a whole.

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

BAYES' THEOREM:

If $A_1, A_2, \dots$ and A_k are mutually exclusive events of which one must occur, then

$$P(A_i | B) = \frac{P(A_i).P(B | A_i)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2) + \dots + P(A_k).P(B | A_k)}$$

If $k = 2$, we obtain:

Bayes' Theorem for two mutually exclusive events A_1 and A_2 :

$$P(A_i | B) = \frac{P(A_i).P(B | A_i)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

Where $i = 1, 2$.

In other words

and

$$P(A_1 | B) = \frac{P(A_1).P(B | A_1)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

$$P(A_2 | B) = \frac{P(A_2).P(B | A_2)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

RANDOM VARIABLE

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).

PROPERTIES OF A DISCRETE PROBABILITY DISTRIBUTION

- (1) $0 \leq P(X_i) \leq 1$ • for each X_i ($i = 1, 2, \dots, 7$)
- (2) $\sum p(X_i) = 1$

the mean is given by:

$$\mu = E(X) = \frac{\sum XP(X)}{\sum p(X)}$$

COMPUTATION OF THE STANDARD DEVIATION

Just as in case of a frequency distribution, we have

$$S = \sqrt{\frac{\sum f(X - \bar{X})^2}{\sum f}}$$

$$= \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2} = \sqrt{\frac{\sum X^2f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2}$$

Similarly, in case of a probability distribution, we have

$$\sigma = \text{S.D.}(X) = \sqrt{\frac{\sum X^2P(X)}{\sum P(X)} - \left[\frac{\sum XP(X)}{\sum P(X)}\right]^2}$$

$$= \sqrt{\sum X^2P(X) - [\sum XP(X)]^2}$$

$$\left(\because \sum P(X) = 1 \right)$$

In the above example

Coefficient of Variation

$$C.V. = \frac{\sigma}{\mu} \times 100$$

DISTRIBUTION FUNCTION

Distribution function, mathematical expression that describes the probability that a system will take on a specific value or set of values. The classic examples are associated with games of chance.

$$F(x) = P(X \leq x)$$

STA301 Short Notes Midterm term

Lec 1 to 22

BY vuonlinehelp.blogspot.com

YOUTUBE LINK:

WEBSITE LINK:

[HTTPS://VUONLINEHELP.BLOGSPOT.COM/](https://vuonlinehelp.blogspot.com/)

IMPORTANT NOTE

You should never copy paste the same answer which we are providing. Please do make some changes in the solution file. If you appreciate our work, please share our website vuonlinehelp.blogspot.com with your friends.

آپ کو کبھی بھی وہی جواب کاپی پیسٹ نہیں کرنا چاہیے جو ہم فراہم کر رہے ہیں۔ براہ کرم حل فائل میں کچھ تبدیلیاں کریں۔ اگر آپ ہمارے کام کی تعریف کرتے ہیں تو براہ کرم ہماری ویب سائٹ vuonlinehelp.blogspot.com کو اپنے دوستوں کے ساتھ شیئر کریں

LECTURE NO. 1 Sta301

More Files Click and VU WhatsApp Group: [Join Now](#)

WHAT IS STATISTICS?

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

Characteristics of Statistics

The important characteristics of Statistics are as follows:

- Statistics are numerically expressed.
- It has an aggregate of facts
- Data are collected in systematic order
- It should be comparable to each other
- Data are collected for a planned purpose

Importance of Statistics

The important functions of statistics are:

- Statistics helps in gathering information about the appropriate quantitative data
- It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
- It provides the exact description and a better understanding
- It helps in designing the effective and proper planning of the statistical inquiry in any field
- It gives valid inferences with the reliability measures about the population parameters from the sample data
- It helps to understand the variability pattern through the quantitative observations

DATA

The word “data” appears in many contexts and frequently is used in ordinary conversation. Although the word carries something of an aura of scientific mystique, its meaning is quite simple and mundane.

Example

In our day to day life, we can collect the following data.

- Number of females per 1000 males in various states of our country.
- Production of wheat in the last 10 years in our country.
- Number of plants in our locality.
- Rainfall in our city in the last 10 years.

- Marks obtained by students.

OBSERVATIONS AND VARIABLES

A **variable** contains all values that measure the same underlying attribute (like height, temperature, duration) across units. An **observation** contains all values measured on the same unit (like a person, or a day, or a city) across attributes.

QUANTITATIVE AND QUALITATIVE VARIABLES

Quantitative data are data about numeric variables (e.g. how many; how much; or how often). **Qualitative data** are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about categorical variables (e.g. what type).

Discrete and continuous variables

A **discrete variable** is a variable that takes on distinct, countable values. A **continuous variable** is a variable that takes on any value within a range, and the number of possible values within that range is infinite.

MEASUREMENT SCALES

long

In Statistics, the variables or numbers are defined and categorized using different scales of measurements.

The four types of scales are:

- Nominal Scale.
- Ordinal Scale.
- Interval Scale.
- Ratio Scale.

Nominal Scale

A **nominal scale** is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.

Ordinal Scale

The **ordinal scale** is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the

“order.” Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

Interval Scale

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

Ratio Scale

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.

ERRORS OF MEASUREMENT

The types ERRORS OF MEASUREMENT are:

- *absolute error*
- *relative error*
- *percentage error*

Absolute Error:

While measuring distances between cities kilometers apart, an error of a few centimeters is negligible and is irrelevant. Consider another case where an error of centimeters when measuring small machine parts is a very significant error. Both the errors are in the order of centimeters but the second error is more severe than the first one.

Relative error

The relative error is defined as the ratio of the absolute error of the measurement to the actual measurement. Using this method we can determine the magnitude of the absolute error in terms of the actual size of the measurement.

Percentage error

Percent error is the difference between estimated value and the actual value in comparison to the actual value and is expressed as a percentage. In other words, the percent error is the relative error multiplied by 100.

LECTURE NO. 2 Sta301

COLLECTION OF DATA

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

PRIMARY AND SECONDARY DATA

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc.

COLLECTION OF PRIMARY DATA

One or more of the following methods are employed to collect primary data:

- Direct Personal Investigation.
- Indirect Investigation.
- Collection through Questionnaires.
- Collection through Enumerators.
- Collection through Local Sources.

DIRECT PERSONAL INVESTIGATION

In direct personal investigation, the investigator obtains the first-hand information from the respondents themselves. He personally visits the respondents to collect the data.

INDIRECT INVESTIGATION

It is a method by which we are gaining the information indirectly by talking to other person who is regarding to have that information.

COLLECTION THROUGH QUESTIONNAIRES

A questionnaire is a list of questions or items used to gather data from respondents about their attitudes, experiences, or opinions. Questionnaires can be used to collect quantitative and/or qualitative information. Questionnaires are commonly used in market research as well as in the social and health sciences.

Collection through Enumerators

In this method, trained enumerators collected the information. They assist the informants in making the entries in the schedules or questionnaires correctly. If the enumerator is

well trained, experienced, and discreet, then you can get the most reliable information through this method.

COLLECTION OF SECONDARY DATA

Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc. Always specific to the researcher's needs.

‘POPULATION’

A **statistical population** is the collection of every member of a group possessing the same basic and defined characteristic, but varying in amount or quality from one member to another.

Frame

The **frame** refers to the list of units (eg, persons, households, businesses, etc) in the survey population. Since the selection of the sample is directly based on this list, the frame is one of the most important tools in the design of a survey.

SAMPLING FRAME

The **sampling frame** is the list from which units are drawn for the sample. The 'list' may be an actual listing of units, as in a phone book from which phone numbers will be sampled, or some other description of the population, such as a map from which areas will be sampled.

ADVANTAGES OF SAMPLING

Advantages of Sampling	Disadvantages of Sampling
Sampling avoid repetition of query for each and every individual	Selection of good samples is difficult
Sampling gives nearest accurate results	Limited knowledge may mislead the result

SAMPLING & NON-SAMPLING ERRORS

Non-sampling error refers to an error that arises from the result of data collection, which causes the data to differ from the true values. It is different from **sampling error**, which is

any difference between the sample values and the universal values that may result from a limited sampling size.

Quota sampling

Quota sampling is a type of non-probability sampling technique. Non-probability sampling focuses on sampling techniques that are based on the judgment of the researcher.

ADVANTAGES OF QUOTA SAMPLING

- There is no need to construct a frame.
- It is a very quick form of investigation.
- Cost reduction.

DISADVANTAGES

- It is a subjective method. One has to choose between objectivity and convenience.
- If random sampling is not employed, it is no longer theoretically possible to evaluate the sampling error.
- Even if the above is not the case, the interviewer may still be making unsuitable selection of sample units.

RANDOM SAMPLING

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population.

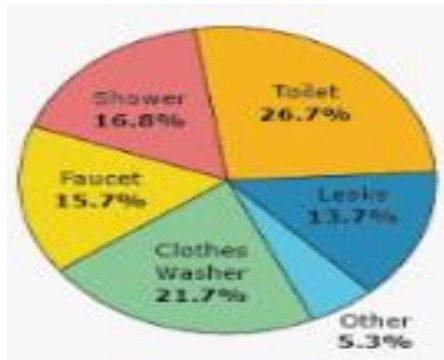
TYPES OF RANDOM SAMPLING

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Cluster Sampling
- Multi-stage Sampling, etc.
-

LECTURE NO. 3 Sta301

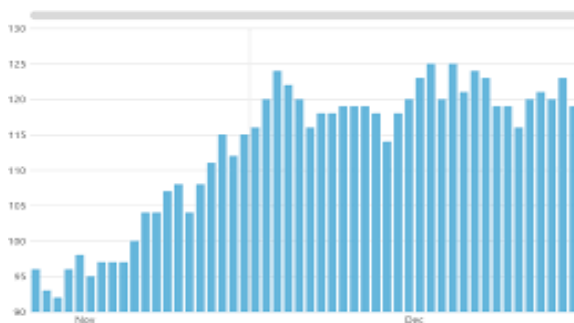
PIE CHART

Pie charts are often used to represent sample data—with data points belonging to a combination of different categories. Each of these categories is represented as a “slice of the pie.”



SIMPLE BAR CHART:

A simple bar chart is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis.



BIVARIATE FREQUENCY TABLE

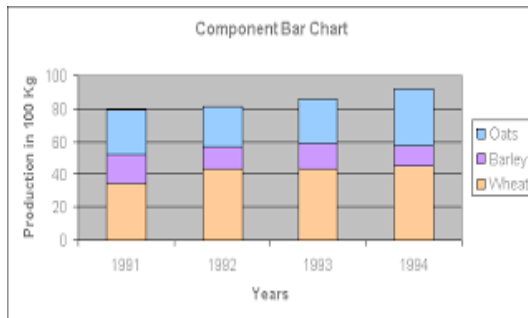
A table that illustrates the relationship between two variables by displaying the distribution of one variable across the categories of a second variable.

Table 3.17
Bivariate Frequency Distribution of Age and Marks

Marks	Age in Years				Total
	16 – 18	18 – 20	20 – 22	22 – 24	
10 – 20	2	1	1	-	4
20 – 30	3	2	3	1	9
30 – 40	3	3	5	6	17
40 – 50	2	2	3	4	11
50 – 60	-	1	2	2	5
60 – 70	-	1	2	1	4
Total	10	10	16	14	50

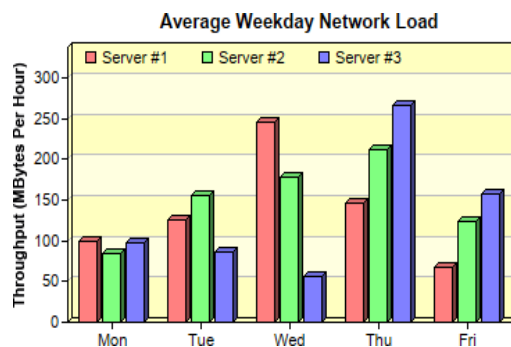
COMPONENT BAR CHART:

A sub-divided or component bar chart is used to represent data in which the total magnitude is divided into different or components.



MULTIPLE BAR CHARTS

A multiple bar graph shows the relationship between different values of data. Each data value is represented by a column in the graph. In a multiple bar graph, multiple data points for each category of data are shown with the addition of columns.



LECTURE NO. 4 Sta301

CONSTRUCTION OF A FREQUENCY DISTRIBUTION

The statistical data that we collect can be presented in the form of a frequency distribution. A frequency distribution refers to summarizing a large data set into a small number of intervals.

Step 1: Sort the data in ascending order

Step 2: Calculate the range of data

Step 3: Decide on the number of intervals in the frequency distribution

Step 4: Determine the intervals.

Step 5: Tally and count the observations under each interval.

$$\text{Relative Frequency} = \frac{\text{Absolute Frequency of Each Interval}}{\text{Total Number of Observations}}$$

HISTOGRAM:

A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

FREQUENCY POLYGON:

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

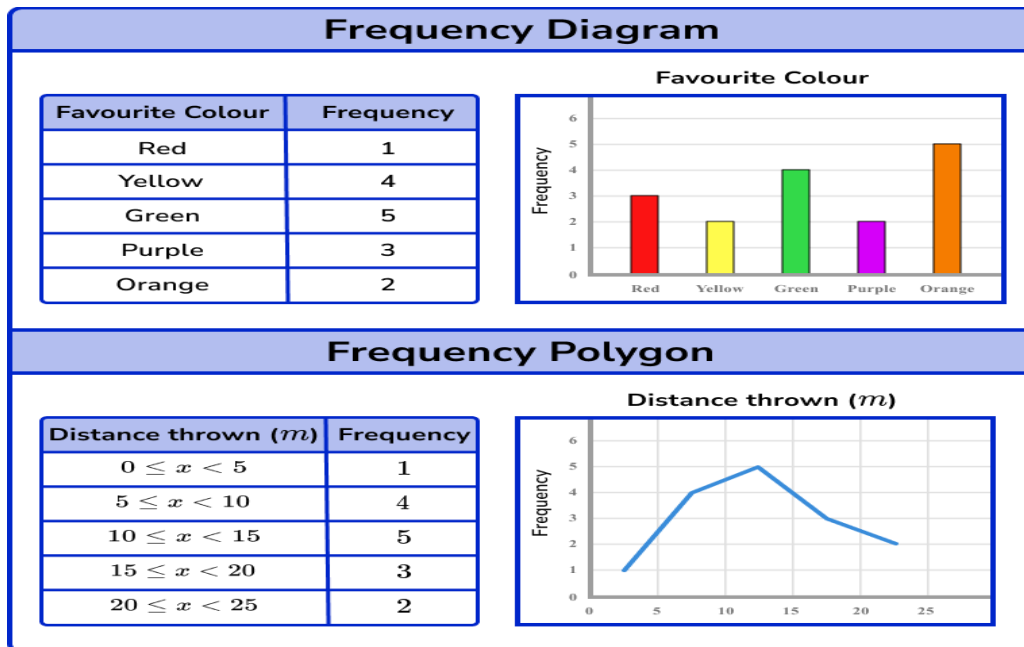
FREQUENCY CURVE:

A Frequency Curve is a smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.

LECTURE NO. 5 Sta301

FREQUENCY POLYGON:

A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.



FREQUENCY CURVE

A **Frequency Curve** is a smooth curve which corresponds to the limiting case of a histogram computed for a frequency distribution of a continuous distribution as the number of data points becomes very large.

VARIOUS TYPES OF FREQUENCY CURVES

- the symmetrical frequency curve
- the moderately skewed frequency curve
- the extremely skewed frequency curve
- the U-shaped frequency curve

CUMULATIVE FREQUENCY DISTRIBUTION

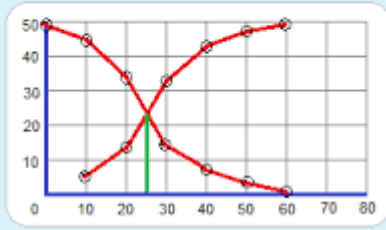
The **cumulative frequency** is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. The last value will always be equal to the total for all observations, since all frequencies will already have been added to the previous total.

CUMULATIVE FREQUENCY POLYGON or OGIVE

The **cumulative frequency polygon** is essentially a line graph drawn on graph paper by plotting actual lower or upper limits of the class intervals on the x-axis and the respective cumulative frequencies of these class intervals on the y-axis.

~ Ogive

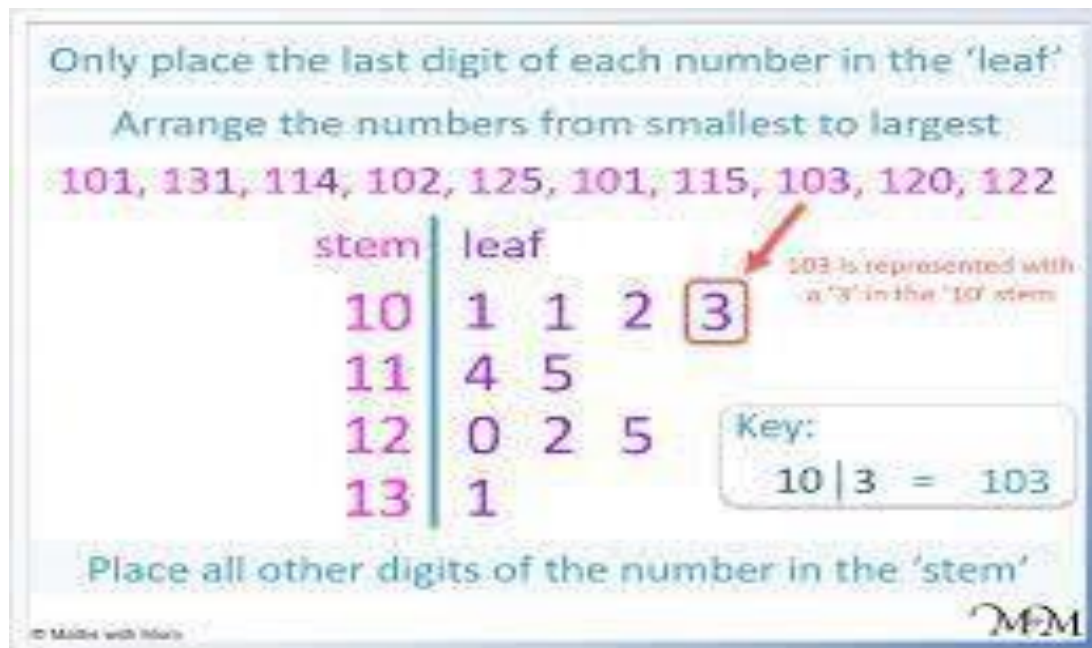
(Cumulative Frequency Curve)



LECTURE NO. 6 Sta301

Stem-and-Leaf Display

A **stem-and-leaf display** or **stem-and-leaf plot** is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution. A stem-and-leaf plot is also called a **stem plot**, but the latter term often refers to another chart type. A simple stem plot may refer to plotting a matrix of y values onto a common x axis, and identifying the common x value with a vertical line, and the individual y values with symbols on the line.



FREQUENCY DISTRIBUTION polygon graph

A **frequency polygon** is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

Class Limits	Class Boundaries	Tally Marks	Frequency
10 - 19	9.5 - 19.5	///	2
20 - 29	19.5 - 29.5	///	3

More Files Click and VU WhatsApp Group: [Join Now](#)

EXAMPLE

Construct a stem-and-leaf display for the data of mean annual death rates per thousand at ages 20-65 given below: 7.5, 8.2, 7.2, 8.9, 7.8, 5.4, 9.4, 9.9, 10.9, 10.8, 7.4, 9.7, 11.6, 12.6, 5.0, 10.2, 9.2, 12.0, 9.9, 7.3, 7.3, 8.4, 10.3, 10.1, 10.0, 11.1, 6.5, 12.5, 7.8, 6.5, 8.7, 9.3, 12.4, 10.6, 9.1, 9.7, 9.3, 6.2, 10.3, 6.6, 7.4, 8.6, 7.7, 9.4, 7.7, 12.8, 8.7, 5.5, 8.6, 9.6, 11.9, 10.4, 7.8, 7.6, 12.1, 4.6, 14.0, 8.1, 11.4, 10.6, 11.6, 10.4, 8.1, 4.6, 6.6, 12.8, 6.8, 7.1, 6.6, 8.8, 8.8, 10.7, 10.8, 6.0, 7.9, 7.3, 9.3, 9.3, 8.9, 10.1, 3.9, 6.0, 6.9, 9.0, 8.8, 9.4, 11.4, 10.9

Using the decimal part in each number as the leaf and the rest of the digits as the stem, we get the ordered stem-and-leaf display shown below:

Stem	Leaf
3	9
4	6 6
5	0 4 5
6	0 0 2 2 5 5 6 6 6 8 9
7	1 3 3 3 4 4 5 6 7 7 8 8 8 9
8	1 1 2 4 6 6 7 7 8 8 8 9 9
9	0 1 2 3 3 3 3 4 4 4 6 7 7 9 9
10	0 1 1 2 3 3 4 4 6 6 7 8 8 9 9
11	1 4 4 6 6 9
12	0 1 4 5 6 8 8
14	0

DESCRIPTION OF VARIABLE DATA:

A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country

of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.

VARIOUS TYPES OF AVERAGES:

There are three different types of average. These are called the mean, the median, and the mode. Each one gives slightly different information.

The Arithmetic, Geometric and Harmonic means are averages that are mathematical in character, and give an indication of the magnitude of the observed values.

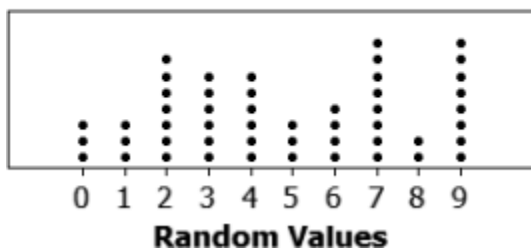
The Median indicates the middle position while the mode provides information about the most frequent value in the distribution or the set of data.

THE MODE: The Mode is defined as that value which occurs most frequently in a set of data i.e. it indicates the most common result.

Dot Plot

a Dot Plot is a graph for displaying the distribution of quantitative variable where each dot represents a value. For whole numbers, if a value occurs more than once, the dots are placed one above the other so that the height of the column of dots represents the frequency for that value.

Dotplot of Random Values



Mode

The mode is the value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all. Other popular measures of central tendency include the mean, or the average of a set, and the median, the middle value in a set.

$$\hat{X} = 1 + \left(\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) \times h$$

Where

l = lower class boundary of the modal class,

f_m = frequency of the modal class,

f_1 = frequency of the class preceding the modal class,

f_2 = frequency of the class following modal class, and

h = length of class interval of the modal class

LECTURE NO. 7 Sta301

THE ARITHMETIC MEAN

The arithmetic mean is a measure of central tendency. It allows us to characterize the center of the frequency distribution of a quantitative variable by considering all of the observations with the same weight afforded to each.

$$\bar{X} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

For simplicity, the above formula can be written as $\bar{X} = \frac{\sum X}{n}$

Example:

Class (Mileage Rating)	Frequency (No. of Cars)
30.0 – 32.9	2
33.0 – 35.9	4
36.0 – 38.9	14
39.0 – 41.9	8
42.0 – 44.9	2
Total	30

Class (Mileage Rating)	Class-mark (Midpoint) X
30.0 – 32.9	31.45
33.0 – 35.9	34.45
36.0 – 38.9	37.45
39.0 – 41.9	40.45
42.0 – 44.9	43.45

Class-mark (Midpoint) X	Frequency f	fX
31.45	2	62.9
34.45	4	137.8
37.45	14	524.3
40.45	8	323.6
43.45	2	86.9
	30	1135.5

Applying the formula

$$\bar{X} = \frac{\sum fX}{\sum f},$$

We obtain

$$\bar{X} = \frac{1135.5}{30} = 37.85$$

MEDIAN

The median is the middle number in an ordered data set. The mean is the sum of all values divided by the total number of values.

MEDIAN

The **MIDDLE** number in a data set

2	4	5	7	12	15	18
			↑			
			Median			

3	4	6	10	13	19
			↙ ↘		
			$\frac{6+10}{2} = \frac{16}{2} = 8$		

52 52 65 73 81 86 89 91 275
Mean = 96 Mode = 52 Median = 81

LECTURE NO. 8 Sta301

Median in Case of a Frequency Distribution of a Continuous Variable:

$$X = l + \frac{h}{f} \left(\frac{n}{2} - c \right)$$

Where

- l = lower class boundary of the median class (i.e. that class for which the cumulative frequency is just in excess of $n/2$).
- h = class interval size of the median class
- f = frequency of the median class
- $n = \sum f$ (the total number of observations)
- c = cumulative frequency of the class preceding the median class

EXAMPLE:

Mileage Rating	No. of Cars	Class Boundaries	Cumulative Frequency
30.0 – 32.9	2	29.95 – 32.95	2
33.0 – 35.9	4	32.95 – 35.95	6
36.0 – 38.9	14	35.95 – 38.95	20
39.0 – 41.9	8	38.95 – 41.95	28
42.0 – 44.9	2	41.95 – 44.95	30

$$\begin{aligned}
 \tilde{X} &= 35.95 + \frac{3}{14}(15 - 6) \\
 &= 35.95 + 1.93 \\
 &= 37.88 \\
 &\simeq 37.9
 \end{aligned}$$

QUARTILES

The quartiles, together with the median, achieve the division of the total area into four equal parts. The first, second and third quartiles are given by the formulae:

4. FIRST QUARTILE

$$Q_1 = l + \frac{h}{f} \left(\frac{n}{4} - c \right)$$

5. second QUARTILE (I.E. MEDIAN)

$$Q_2 = l + \frac{h}{f} \left(\frac{2n}{4} - c \right) = l + \frac{h}{f} (n/2 - c)$$

6. THIRD QUARTILE

$$Q_3 = l + \frac{h}{f} \left(\frac{3n}{4} - c \right)$$

DECILES & PERCENTILES

The deciles and the percentiles give the division of the total area into 10 and 100 equal parts respectively. The formula for the first decile is:

$$D_1 = 1 + \frac{h}{f} \left(\frac{n}{10} - c \right)$$

$$D_2 = 1 + \frac{h}{f} \left(\frac{2n}{10} - c \right)$$

$$D_3 = 1 + \frac{h}{f} \left(\frac{3n}{10} - c \right)$$

It is easily seen that the 5th decile is the same quantity as the median. The formula for the first percentile is:

$$P_1 = 1 + \frac{h}{f} \left(\frac{n}{100} - c \right)$$

$$P_2 = 1 + \frac{h}{f} \left(\frac{2n}{100} - c \right)$$

$$P_3 = 1 + \frac{h}{f} \left(\frac{3n}{100} - c \right)$$

LECTURE NO. 9 Sta301

GEOMETRIC MEAN

In Mathematics, the Geometric Mean (GM) is the average value or mean which signifies the central tendency of the set of numbers by finding the product of their values.

Basically, we multiply the numbers altogether and take the nth root of the multiplied numbers, where n is the total number of data values.

$$G = \sqrt[n]{X_1 X_2 \dots X_n}$$

$$\log G = \frac{1}{n} [\log X_1 + \log X_2 + \dots + \log X_n]$$

$$= \frac{\sum \log X}{n}$$

$$G = \text{antilog} \left[\frac{\sum \log X}{n} \right]$$

EXAMPLE:

Find the geometric mean of numbers:

45, 32, 37, 46, 39, 36, 41, 48, 36

Solution:

We need to compute the numerical value of

$$= \sqrt[9]{45 \times 32 \times 37 \times 46 \times 39 \times 36 \times 41 \times 48 \times 36}$$

But, obviously, it is a bit cumbersome to find the ninth root of a quantity. So we make use of logarithms, as shown below:

X	log X
45	1.6532
32	1.5052
37	1.5682
46	1.6628
39	1.5911
36	1.5563
41	1.6128
48	1.6812
36	1.5563
	14.3870

$$\log G = \frac{\sum \log X}{n}$$

$$= \frac{14.3870}{9} = 1.5986$$

$$\text{Hence } G = \text{anti log } 1.5986$$

$$= 39.68$$

GEOMETRIC MEAN FOR GROUPED DATA

In case of a frequency distribution having k classes with midpoints $X_1, X_2, \dots, X_k$ and the corresponding frequencies $f_1, f_2, \dots, f_k$ (such that $\sum f_i = n$), the geometric mean is given by

$$G = \sqrt[n]{X_1^{f_1} X_2^{f_2} \dots X_k^{f_k}}$$

$$\log G = \frac{1}{n} \left[f_1 \log X_1 + f_2 \log X_2 + \dots + f_k \log X_k \right]$$

$$G = \frac{\sum f \log X}{n}$$

$$G = \text{anti log} \left[\frac{\sum f \log X}{n} \right]$$

RELATION BETWEEN ARITHMETIC, GEOMETRIC AND HARMONIC MEANS

MID-RANGE

If there are n observations with x_0 and x_m as their smallest and largest observations respectively, then their mid-range is defined as

$$\text{mid - range} = \frac{x_0 + x_m}{2}$$

MID-QUARTILE RANGE

If $x_1, x_2 \dots x_n$ are n observations with Q_1 and Q_3 as their first and third quartiles respectively, then their mid-quartile range is defined as

$$\text{mid - quartile range} = \frac{Q_1 + Q_3}{2}$$

LECTURE NO. 10 Sta301

ABSOLUTE VERSUS RELATIVE

There are two types of measurements of dispersion: absolute and relative. An absolute measure of dispersion is one that measures the dispersion in terms of the same units or in the square of units, as the units of the data. For example, if the units of the data are rupees, meters, kilograms, etc., the units of the measures of dispersion will also be rupees, meters, kilograms, etc. On the other hand, relative measure of dispersion is one that is expressed in the form of a ratio, co-efficient of percentage and is independent of the units of measurement.

RANGE

The range is defined as the difference between the two extreme values of a data-set, i.e. $R = X_m - X_0$ where X_m represents the highest value and X_0 the lowest. Evidently, the calculation of the range is a simple question of MENTAL arithmetic.

Disadvantages

3. It *ignores* all the **INFORMATION** available from the intermediate observations.
4. It might give a **MISLEADING** picture of the spread in the data.

From **THIS** point of view, it is an unsatisfactory measure of dispersion. However, it is **APPROPRIATELY** used in statistical quality control charts of manufactured products, daily temperatures, stock prices, etc.

$$\begin{aligned} & \frac{(Midrange - X_0) + (X_m - Midrange)}{2} \times 2 \\ &= \frac{Midrange - X_0 + X_m - Midrange}{2} \\ &= \frac{X_m - X_0}{2} \end{aligned}$$

The range is an *absolute* measure of dispersion. Its *relative* measure is known as the *Coefficient Of Dispersion*, and is defined by the relation given below:

$$= \frac{\frac{1}{2} (Range)}{Mid - Range}$$

$$\frac{\frac{X_m - X_0}{2}}{\frac{X_m + X_0}{2}}$$

QUARTILE DEVIATION

The quartile deviation is defined as half of the difference between the third and first quartiles i.e.

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

COEFFICIENT OF QUARTILE DEVIATION

$$\begin{aligned} &= \frac{\text{Quartile Deviation}}{\text{Mid - Quartile Range}} \\ &= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}, \end{aligned}$$

MEAN DEVIATION

Mean deviation is a measure of dispersion, which is known as the average deviation. Mean deviation can be computed from the mean or median. Mean deviation is the arithmetic deviation of different items of central tendency. It may be the mean or the median.

$$\text{M.D.} = \frac{\sum |d|}{n}$$

LECTURE NO. 11 Sta301

MEAN DEVIATION

The mean deviation is the mean of the absolute deviations of the observations or values from a suitable average. This suitable average may be the mean, median or mode. We also know it as the mean absolute deviation.

The basic formula to calculate mean deviation for a given data set is as follows:

$$\text{Mean Deviation} = \frac{\sum |X - \bar{X}|}{N}$$

Where,

X = denotes each value in the data set

$\bar{X}$ = denotes the mean value of the data set

N = total number of data values

$||$ = represents absolute value, i.e. it ignores the sign

MEAN DEVIATION FOR GROUPED DATA

Ungrouped data is data that has not been sorted or categorized into groups and is still in its raw form. Generally, ungrouped data includes individual data series. The formula to calculate mean deviation for ungrouped data is as follows:

$$\text{Mean Deviation} = \frac{\sum_1^n |x_i - \bar{x}|}{n}$$

Where,

$x_i = i^{\text{th}}$ observation

$\bar{x}$ = central point of the data (mean, median or mode)

n = number of observations

Variance

The sum of squares of the deviations of the X from the mean, divided by the no. of values.

$$\text{STANDARD DEVIATION} = \frac{\sum (x - \bar{x})^2}{n}$$

Variance is equal to the average squared deviations from the mean, while standard deviation is the number's square root. Also, the standard deviation is a square root of variance.

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

STANDARD DEVIATION IN CASE OF GROUPED DATA

$$S = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$$

SHORT CUT FORMULA OF THE STANDARD DEVIATION IN CASE OF GROUPED DATA

$$S = \sqrt{\left\{ \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n} \right)^2 \right\}}$$

COEFFICIENT OF VARIATION

The coefficient of variation (CV) is the ratio of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population. The higher the CV, the greater the dispersion.

$$\text{C.V.} = \frac{S}{\bar{X}} \times 100$$

LECTURE NO. 12 Sta301

CHEBYCHEV'S THEOREM

Chebyshev's Theorem estimates the minimum proportion of observations that fall within a specified number of standard deviations from the mean. This theorem applies to a broad range of probability distributions. Chebyshev's Theorem is also known as Chebyshev's Inequality.

Here are two forms of the equation. One determines how close to the mean the data lie and the other calculate how far away from the mean they fall:

Maximum proportion of observations that are more than k standard
deviations from the mean

$$\frac{1}{k^2}$$

Minimum proportion of observations that are within k standard deviations of the mean

$$1 - \frac{1}{k^2}$$

Where k equals the number of standard deviations in which you are interested. K must be greater than 1.

Five-number summaries

A **five-number summary** is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q_1), median value (Q_2), upper quartile (Q_3), maximum value.

LECTURE NO. 13 Sta301

BOX AND WHISKER PLOT

A **box and whisker plot** is **defined** as a graphical method of displaying variation in a set of data. In most cases, a histogram analysis provides a sufficient display, but a box and whisker plot can provide additional detail while allowing multiple sets of data to be displayed in the same graph.

The procedure to develop a box and whisker plot comes from the five statistics below. You can also download the box and whisker plot template.

6. **Minimum value:** The smallest value in the data set
7. **Second quartile:** The value below which the lower 25% of the data are contained
8. **Median value:** The middle number in a range of numbers
9. **Third quartile:** The value above which the upper 25% of the data are contained
10. **Maximum value:** The largest value in the data set

INTERPRETATION OF THE BOX AND WHISKER PLOT

In a box and whisker plot: The left and right sides of the box are the lower and upper quartiles. The box covers the interquartile interval, where 50% of the data is found. The vertical line that split the box in two is the median.

COEFFICIENT OF SKEWNESS

The coefficient of skewness measures the skewness of a distribution. It is based on the notion of the moment of the distribution. This coefficient is one of the measures of skewness.

$$\text{skewness} = \frac{\bar{x} - \text{mode}}{S},$$

LECTURE NO. 14 Sta301

PEARSON'S COEFFICIENT OF SKEWNESS:

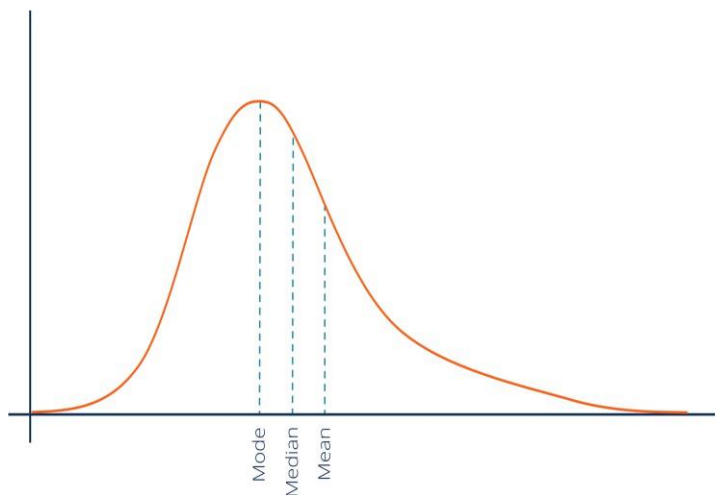
pearson's coefficient of skewness (second method) is calculated by multiplying the difference between the mean and median, multiplied by three. The result is divided by the standard deviation. You can use the Excel functions AVERAGE, MEDIAN and STDEV.P to get a value for this measure.

In Power BI, this is implemented by the following formula:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

POSITIVE SKEWNESS

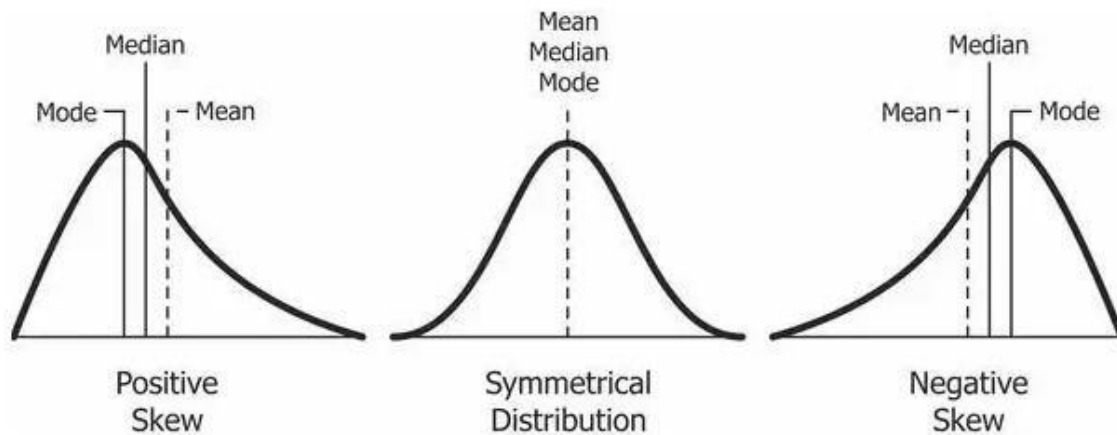
In statistics, a positively skewed (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer. The positively skewed distribution is the direct opposite of the negatively skewed distribution.



Mean > Median > Mode

NEGATIVE SKEWNESS

Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.

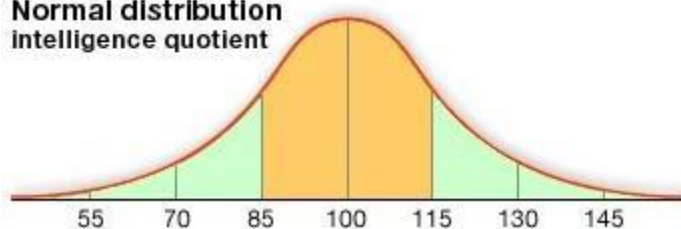


$3 * (\text{Mean} - \text{Median}) / \text{Standard Deviation}$.

Relative

A relative frequency distribution shows the proportion of the total number of observations associated with each value or class of values and is related to a probability distribution, which is extensively used in statistics.

Normal distribution intelligence quotient



Hint: To find the Bowley's coefficient of skewness, we first need to calculate three quartiles at $\frac{N}{4}$,

$\frac{N}{2}$ and $\frac{3N}{4}$ respectively. Then, we can calculate the Bowley's coefficient of skewness using the

$$\text{formula } SK_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}.$$

KURTOSIS

Kurtosis is a measure of the tailedness of a distribution. Tailedness is how often outliers occur. **Excess kurtosis** is the tailedness of a distribution relative to a normal distribution.

- Distributions with medium kurtosis (medium tails) are **mesokurtic**.
- Distributions with low kurtosis (thin tails) are **platykurtic**.
- Distributions with high kurtosis (fat tails) are **leptokurtic**.

Tails are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, tails represent how often **outliers** occur.

Leptokurtic

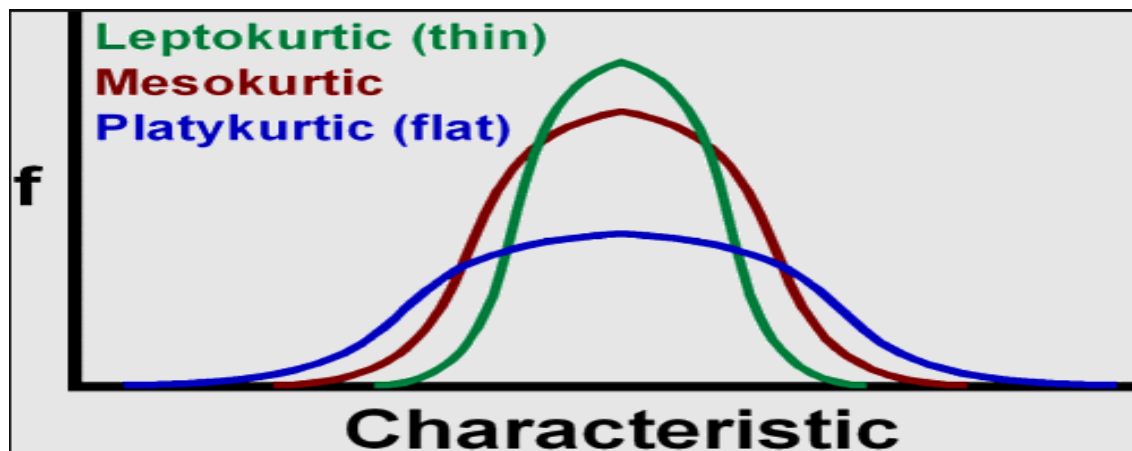
Leptokurtic distributions are statistical distributions with kurtosis greater than three. It can be described as having a wider or flatter shape with fatter tails resulting in a greater chance of extreme positive or negative events. It is one of three major categories found in kurtosis analysis.

Platykurtic

The term "platykurtic" refers to a statistical distribution in which the excess kurtosis value is **negative**. For this reason, a platykurtic distribution will have thinner tails than a normal distribution will, resulting in fewer extreme positive or negative events.

Mesokurtic

Mesokurtic is a statistical term used to describe the outlier characteristic of a probability distribution that is **close to zero**. Mesokurtic distributions are similar to normal distributions, in which extreme or outlier events are very unlikely.



MOMENTS

The third moment is called skewness, and the fourth moment is known as kurtosis. The third moment measures the asymmetry of distribution while the fourth moment measures how heavy the tail values are. Physicists generally use the higher-order moments in applications of physics.

$$\frac{1}{n} \sum (x_i - \bar{x})^1 = \frac{1}{n} \sum (x_i - \bar{x})$$

Is called the first sample moment about the mean, and is denoted by m_1 . Similarly, the quantity.

$$\frac{1}{n} \sum (x_i - \bar{x})^2$$

Is called the second sample moment about the mean, and is denoted by m_2 .

$$m_r = \frac{1}{n} \sum (x_i - \bar{x})^r$$

Moments about the mean are also called the central moments or the mean moments.

$$m'_r = \frac{1}{n} \sum (x_i - \alpha)^r \quad \text{for sample data}$$

For $r = 1$, we have

$$m_1 = \frac{1}{n} \sum (x_i - \bar{x}) = \frac{\sum x_i}{n} - \bar{x} = \bar{x} - \bar{x} = 0,$$

and

$$m'_1 = \frac{1}{n} \sum (x_i - \alpha) = \frac{\sum x_i}{n} - \alpha = \bar{x} - \alpha.$$

Putting $r = 2$ in the relation for mean moments, we see that

$$m_2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

if we put $\alpha = 0$, we obtain

$$m'_r = \frac{1}{n} \sum (x_i - \alpha)^r$$

$$m'_r = \frac{1}{n} \sum x_i^r$$

SHEPPARD'S CORRECTIONS:

In statistics, **Sheppard's corrections** are approximate corrections to estimates of moments computed from binned data. The concept is named after William Fleetwood Sheppard.

Let m_k be the measured k^{th} moment, $\hat{\mu}_k$ the corresponding corrected moment, and c the breadth of the class interval. No correction is necessary for the mean (first moment about zero). The first few measured and corrected moments about the mean are then related as follows:

$$\hat{\mu}_2 = m_2 - \frac{1}{12}c^2$$

$$\hat{\mu}_3 = m_3$$

$$\hat{\mu}_4 = m_4 - \frac{1}{2}m_2c^2 + \frac{7}{240}c^4.$$

The reason for doing so is that, in many situations, it is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the mean-moments using the relationships that I am now going to convey to you. The equations are:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2;$$

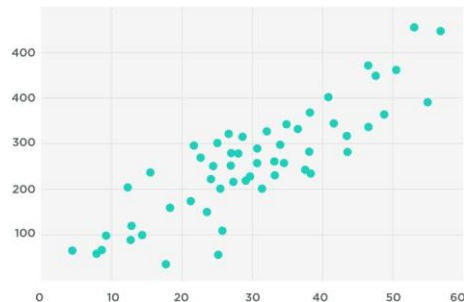
$$m_3 = m'_3 - 3 m'_2 m'_1 + 2 (m'_1)^3, \text{ and}$$

$$m_4 = m'_4 - 4 m'_3 m'_1 + 6 m'_2 (m'_1)^2 - 3 (m'_1)^4$$

LECTURE NO. 15 Sta301

Scatter plot

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.



Scatter formula

Step 1: Take any two points on the line; preferably the points that are at the intersection of the grid. We can take any points on the line but it gets easier with the calculations if we take a point at the intersection of the grid. Find the coordinates of the points we identified.

Step 2: Find the slope of the line using the points.

Find the slope of the line by substituting the value of the coordinates in the equation

$slope(m) = \frac{y_2 - y_1}{x_2 - x_1}$, where (x_1, y_1) are the coordinates of one point and (x_2, y_2) are the coordinates of second point.

Step 3: Find the y-intercept. To find the y- intercept (b), take any point identified in step 1 and substitute the value of the x-coordinate and the y-coordinate of the point and the value of slope (m) found in step 2 in the equation $y = mx + b$.

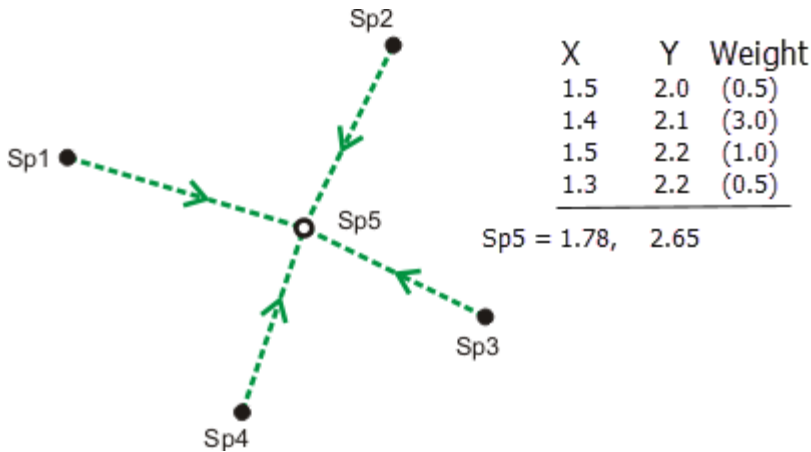
Step 4: Find the equation of a given line of best fit for the given scatter plot.

The equation of a line of fit is $y = mx + b$, where m is the slope found in step 2 and b is the y-intercept found in step 3.

Step 5: To make predictions, take a look at the given word problem and substitute the given value in the equation of the line of fit we found in the previous step to predict the unknown value asked.

THE PRINCIPLE OF LEAST SQUARES

A least-squares adjustment uses statistical analysis to estimate the most likely coordinates for connected points in a measurement in a network. In ArcGIS Pro, a least-squares adjustment can be run on a parcel fabric using the following tools: Analyze Parcels By Least Squares Adjustment tool.



$$\left. \begin{aligned} \sum Y &= na + b \sum X \\ \sum XY &= a \sum X + b \sum X^2 \end{aligned} \right\}$$

STANDARD ERROR OF ESTIMATE

The standard error of the estimate is a measure of the average deviation of the errors, the difference between the $\hat{y}$ -values predicted by the multiple regression model and the y -values in the sample. The standard error of the estimate for the regression model is the standard deviation of the errors/residuals.

$$s_{y.x} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}}$$

$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{n - 2}}$$

Step 1: Note the number of measurements (n) and determine the sample mean (μ). It is the average of all the measurements.

Step 2: Determine how much each measurement varies from the mean.

Step 3: Square all the deviations determined in step 2 and add altogether: $\sum (x_i - \mu)^2$

Step 4: Divide the sum from step 3 by one less than the total number of measurements ($n-1$).

Step 5: Take the square root of the obtained number, which is the standard deviation (σ).

Step 6: Finally, divide the standard deviation obtained by the square root of the number of measurements (n) to get the standard error of your estimate.

Go through the example given below to understand the method of calculating standard error.

INTERPRETATION OF $S_{y.x}$

$S_{y.x}$ is computed in a very similar way, but the denominator is $n-K$, where K is the number of parameters fit by regression. The value $n-K$ is the number of degrees of freedom of the regression. If you only fit one parameter, then the RMSE and $S_{y.x}$ are the same.

CORRELATION

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

PEARSON'S COEFFICIENT OF CORRELATION

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

where, covariance of X and Y is defined as
$$r = \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}}$$

$$Cov(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$

SHORT CUT FORMULA FOR THE PEARSON'S COEFFICIENT OF CORRELATION

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

LECTURE NO. 16 Sta301

“SET”

A set is the mathematical model for a collection of different things a set contains elements or members, which can be mathematical objects of any kind: numbers, symbols, points in space, lines, other geometrical shapes, variables, or even other sets. The set with no element is the empty set; a set with a single element is a singleton. A set may have a finite number of elements or be an infinite set. Two sets are equal if they have precisely the same elements.

$$A = \{a, b, c, d\} \text{ or} \\ B = \{1, 2, 3, 7\}$$

Type of sets

Empty Sets

The set, which has no elements, is also called a null set or void set. It is denoted by $\{\}$. Below are the two examples of the empty set.

Singleton Sets

The set which has just one element is named a singleton set. For example, Set $A = \{8\}$ is a singleton set.

Finite and Infinite Sets

A set that has a finite number of elements is known as a finite set, whereas the set whose elements can't be estimated, but has some figure or number, which is large to precise in a set, is known as infinite set.

For example, set $A = \{3, 4, 5, 6, 7\}$ is a finite set, as it has a finite number of elements.

Equal Sets

If every element of set A is also the elements of set B and if every element of set B is also the elements of set A, then sets A and B are called equal sets. It means set A and set B have equivalent elements and that we can denote it as:

$$A = B$$

For example, let $A = \{3, 4, 5, 6\}$ and $B = \{6, 5, 4, 3\}$, then $A = B$

Subsets

A set S is said to be a subset of set T if the elements of set S belong to set T, or you can say each element of set S is present in set T. Subset of a set is denoted by the symbol ($\subset$) and written as $S \subset T$.

We can also write the subset notation as:

$$S \subset T \text{ if } p \in S \Rightarrow p \in T$$

Power Sets

The set of all subsets is known as power sets. We know the empty set is a subset of all sets, and each set is a subset of itself. Taking an example of set $X = \{2,3\}$. From the above-given statements, we can write,

$\{\}$ is a subset of $\{2,3\}$

$\{2\}$ is a subset of $\{2,3\}$

$\{3\}$ is a subset of $\{2,3\}$

$\{2,3\}$ is also a subset of $\{2,3\}$

Therefore, power set of $X = \{2,3\}$,

$P(X) = \{\{\}, \{2\}, \{3\}, \{2,3\}\}$

Universal Sets

A set that contains all the elements of other sets is called a universal set. Generally, it is represented as 'U.'

For example, set $A = \{1,2,3\}$, set $B = \{3,4,5,6\}$, and $C = \{5,6,7,8,9\}$.

Disjoint Sets

If two sets X and Y do not have any common elements, and their intersection results in zero(0), then set X and Y are called disjoint sets. It can be represented as, $X \cap Y = 0$.

OPERATIONS ON SETS

Union of Sets

The union of two sets consists of all their elements. It is denoted by (U) .

For example, set $A = \{2,3,7\}$ and set $B = \{4,5,8\}$.

Then the union of set A and set B will be:

$A \cup B = \{2,3,7,4,5,8\}$

Intersection of Sets

The set of all elements, which are common to all the given sets, gives an intersection of sets. It is denoted by $\cap$.

For example, set $A = \{2,3,7\}$ and set $B = \{2,4,9\}$.

So, $A \cap B = \{2\}$

Difference of Sets

The difference between set S and set T is such that it has only those elements which are in the set S and not in the set T . $S - T = \{p : p \in S \text{ and } p \notin T\}$

Similarly, $T - S = \{p : p \in T \text{ and } p \notin S\}$.

Complement of a Set

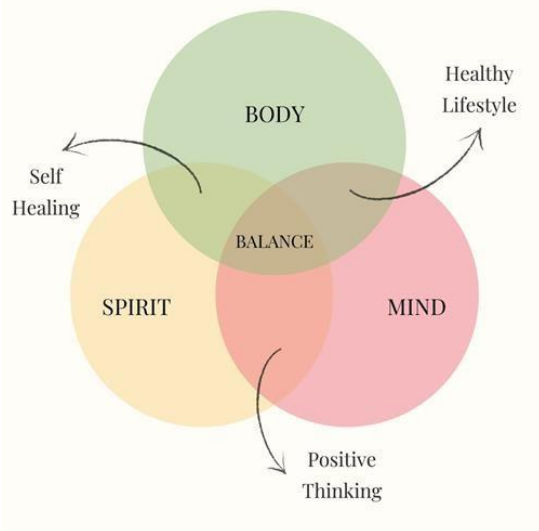
Let U be the universal set and let $A \subset U$. Then, the complement of A , denoted by A' or $(U - A)$, is defined as

$A' = \{x \in U : x \notin A\}$

$X A' \times A$

VENN DIAGRAM

A Venn diagram is an illustration that uses circles to show the relationships among things or finite groups of things. Circles that overlap have a commonality while circles that do not overlap do not share those traits. Venn diagrams help to visually represent the similarities and differences between two concepts.



COMPLEMENTATION:

The **complement of set A** is defined as a set that contains the elements present in the universal set but not in set A. For example, Set $U = \{2,4,6,8,10,12\}$ and set $A = \{4,6,8\}$, then the complement of set A, $A' = \{2,10,12\}$.

Commutative laws

$$A \cup B = B \cup A \text{ and}$$

$$A \cap B = B \cap A$$

Associative laws

$$(A \cup B) \cup C = A \cup (B \cup C)$$

and

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive laws

$$A \cap (B \cup C)$$

$$= (A \cap B) \cup (A \cap C)$$

and $A \cup (B \cap C)$

$$= (A \cup B) \cap (A \cup C)$$

Idempotent laws

$$A \cup A = A \text{ and}$$

$$A \cap A = A$$

Identity laws

$$A \cup S = S,$$

$$A \cap S = A,$$

$$A \cup \phi = A, \text{ and}$$

$$A \cap \phi = \phi.$$

Complementation laws

$$A \cup \bar{A} = S,$$

$$A \cap \bar{A} = \phi,$$

$$(\bar{\bar{A}}) = A,$$

$$\bar{\bar{S}} = \phi, \text{ and}$$

$$\bar{\phi} = S$$

De Morgan's laws

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B},$$

$$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$$

PARTITION OF SETS

A partition of a set S is a sub-division of the set into non-empty subsets that are disjoint and exhaustive, i.e. their union is the set S itself.

This implies that:

- i) $A_i \cap A_j = \phi$, where $i \neq j$;
- ii) $A_1 \cap A_2 \cup \dots \cup A_n = S$.

The subsets in a partition are called cells.

CLASS OF SETS

A set of sets is called a class. For example, in a set of lines, each line is a set of points.

POWER SET

The class of ALL subsets of a set A is called the Power Set of A and is denoted by $P(A)$.

For example, if $A = \{H, T\}$, then $P(A) = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

CARTESIAN PRODUCT OF SETS

The Cartesian product of sets A and B , denoted by $A \times B$, (read as “A cross B”), is a set that contains all ordered pairs (x, y) , where x belongs to A and y belongs to B .

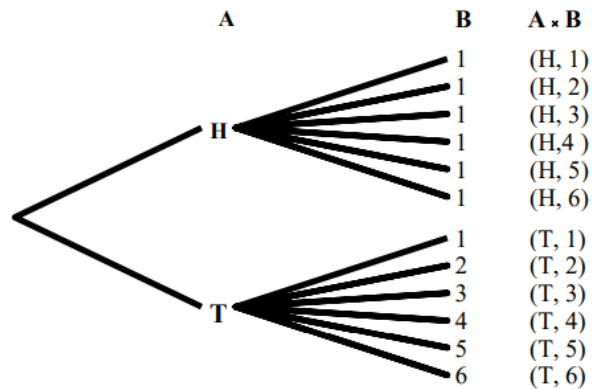
Symbolically, we write

$$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}$$

TREE DIAGRAM

In probability theory, a **tree diagram** may be used to represent a probability space. Tree diagrams may represent a series of independent events (such as a set of coin flips) or conditional probabilities (such as drawing cards from a deck, without replacing the cards). Each node on the diagram represents an event and is associated with the probability of that event. The root node represents the certain event and therefore has probability. Each set of sibling nodes represents an exclusive and exhaustive partition of the parent event.

Tree Diagram



LECTURE NO. 17 Sta301

COUNTING RULES

The Fundamental Counting Principle (also called the counting rule) is a way to figure out the number of outcomes in a probability problem. Basically, you multiply the events together to get the total number of outcomes.

RULE OF PERMUTATION

The special permutation rule states that anything permute itself is equivalent to itself factorial.

Example: $(3)3 = 3!$

$${}^nP_r = n(n-1)(n-2) \dots (n-r+1)$$

$$= \frac{n!}{(n-r)!}$$

$$\text{is } P = \frac{n!}{n_1! n_2! \dots n_k!}$$

(where $\sum n_i = n$)

RULE OF COMBINATION

Dempster's rule of combination is a rule for combining two or more belief functions; when the belief functions combined are based on distinct or “independent” sources of evidence, the rule corresponds intuitively to the pooling of evidence.

and is given by

$${}^nC_r \text{ or } \binom{n}{r},$$
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

It should be noted that

$${}^nP_r = r! \binom{n}{r}$$

RANDOM EXPERIMENT

The Random Experiment: A Precise Definition of a Random Situation. A random experiment is any well-defined procedure that produces an observable outcome that could not be perfectly predicted in advance. A random experiment must be well defined to eliminate any vagueness or surprise.

PROPERTIES OF A RANDOM EXPERIMENT

A random experiment is a process characterized by the following properties:

- (iv) It is performed according to some set of rules,
- (v) It can be repeated arbitrarily often,
- (vi) The result of each performance depends on chance and cannot be predicted uniquely.

Sample space

A **sample space** is the set of all possible outcomes of a statistical experiment, and it is sometimes referred to as a probability space. And outcomes are observations of the experiment, and they are sometimes referred to as sample points.

Events

In probability, **events** are the outcomes of an experiment. The probability of an event is the measure of the chance that the event will occur as a result of an experiment.

SIMPLE & COMPOUND EVENTS

Simple events are events where one experiment happens at a time and it will have a single outcome such as tossing a coin. **Compound events** are events where there is more than one possible outcome such as roll a five using a 6-sided die.

OCCURRENCE OF AN EVENT

An event A is said to occur if and only if the outcome of the experiment corresponds to some element of A.

COMPLEMENTARY EVENT

In probability theory, the complement of any event A is the event [not A], i.e. the event that A does not occur. The event A and its complement [not A] are mutually exclusive and exhaustive.

MUTUALLY EXCLUSIVE EVENTS

In statistics and probability theory, two events are mutually exclusive if they cannot occur at the same time. The simplest example of mutually exclusive events is a coin toss. A tossed coin outcome can be either head or tails, but both outcomes cannot occur simultaneously.

EXHAUSTIVE EVENTS

A set of events are called exhaustive events if at least one of them necessarily occurs whenever the experiment is performed. Also, the union of all these events constitutes the sample space of that experiment.

PARTITION OF THE SAMPLE SPACE

A collection of sets $B_1, B_2, \dots, B_n$ is said to partition the sample space if the sets

- (iii) are mutually disjoint and
- (iv) Have as union the entire sample space.

A simple example of a partition is given by a set B, together with its complement B^c .

EQUALLY LIKELY EVENTS

Equally likely events are events that have the same theoretical probability (or likelihood) of occurring. Each numeral on a die is equally likely to occur when the die is tossed.

LECTURE NO. 18 Sta301

MUTUALLY EXCLUSIVE EVENTS

In statistics and probability theory, two events are mutually exclusive if they cannot occur at the same time. The simplest example of mutually exclusive events is a coin toss. A tossed coin outcome can be either head or tails, but both outcomes cannot occur simultaneously.

EXHAUSTIVE EVENTS

Set of events are called exhaustive events if at least one of them necessarily occurs whenever the experiment is performed. Also, the union of all these events constitutes the sample space of that experiment.

PARTITION OF THE SAMPLE SPACE

Partitions: A collection of sets $B_1, B_2, \dots, B_n$ is said to partition the sample space if the sets (i) are mutually disjoint and (ii) have as union the entire sample space. A simple example of a partition is given by a set B , together with its complement B .

EQUALLY LIKELY EVENTS

Equally likely events are events that have the same theoretical probability (or likelihood) of occurring. Each numeral on a die is equally likely to occur when the die is tossed.

SUBJECTIVE OR PERSONALISTIC PROBABILITY

Subjective probability is a type of probability derived from an individual's personal judgment or own experience about whether a specific outcome is likely to occur. It contains no formal calculations and only reflects the subject's opinions and past experience.

THE CLASSICAL OR 'A PRIORI' DEFINITION OF PROBABILITY

A priori probability refers to the likelihood of an event occurring when there is a finite amount of outcomes and each is equally likely to occur. The outcomes in a priori probability are not influenced by the prior outcome.

$$P(A) = \frac{m}{n}$$
$$= \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

LECTURE NO. 19 Sta301

STATISTICAL PROBABILITY

The probability is the measure of the likelihood of an event to happen. It measures the certainty of the event. The formula for probability is given by;

$P(E) = \text{Number of Favourable Outcomes} / \text{Number of total outcomes.}$

$P(E) = n(E)/n(S)$

Here,

$n(E)$ = Number of event favourable to event E

$n(S)$ = Total number of outcomes

LAW OF COMPLEMENTATION

The union of a set A and its complement A' gives the universal set U of which, A and A' are a subset. $A \cup A' = U$. Also, the intersection of a set A and its complement A' gives the empty set $\emptyset$. $A \cap A' = \emptyset$

For Example: If $U = \{1, 2, 3, 4, 5\}$ and $A = \{1, 2, 3\}$ then $A' = \{4, 5\}$.

$$P(\overline{A}) = 1 - P(A).$$

ADDITION LAW

The addition rule states the probability of two events is the sum of the probabilities of two events that will happen minus the probability of both the events that will happen. Mathematically, the addition rule of probability is expressed as:
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

LECTURE NO. 20 Sta301

ADDITION LAW

The addition rule states the probability of two events is the sum of the probabilities of two events that will happen minus the probability of both the events that will happen. Mathematically, the addition rule of probability is expressed as:
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

COROLLARY-1

If A and B are mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B)$$

(Since $A \cap B$ is an impossible event, hence $P(A \cap B) = 0$.)

COROLLARY-2

If $A_1, A_2, \dots, A_k$ are k mutually exclusive events, then the probability that one of them occurs, is the sum of the probabilities of the separate events, i.e.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

Conditional probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

Where $P(B) > 0$
 (If $P(B) = 0$, the conditional probability $P(A/B)$ remains undefined.)
 Similarly

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

MULTIPLICATION LAW

(The probability of A given B equals the probability of A and B divided by the probability of B.)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

LECTURE NO. 21 Sta301

INDEPENDENT EVENTS

Independent events are those events whose occurrence is not dependent on any other event. For example, if we flip a coin in the air and get the outcome as Head, then again if we flip the coin but this time we get the outcome as Tail. In both cases, the occurrence of both events is independent of each other.

If A and B are independent events, then $P(A | B) = P(A)$

Using the Multiplication rule of probability, $P(A \cap B) = P(B) \cdot P(A | B)$

$$P(A \cap B) = P(B) \cdot P(A)$$

LECTURE NO. 22 Sta301

BAYES' THEOREM

In probability theory and statistics, **Bayes' theorem** (alternatively **Bayes' law** or **Bayes' rule**), named after Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

For example, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on their age) than simply assuming that the individual is typical of the population as a whole.

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

BAYES' THEOREM:

If $A_1, A_2, \dots$ and A_k are mutually exclusive events of which one must occur, then

$$P(A_i | B) = \frac{P(A_i).P(B | A_i)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2) + \dots + P(A_k).P(B | A_k)}$$

If $k = 2$, we obtain:

Bayes' Theorem for two mutually exclusive events A_1 and A_2 :

$$P(A_i | B) = \frac{P(A_i).P(B | A_i)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

Where $i = 1, 2$.

In other words

$$\text{and } P(A_1 | B) = \frac{P(A_1).P(B | A_1)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

$$P(A_2 | B) = \frac{P(A_2).P(B | A_2)}{P(A_1).P(B | A_1) + P(A_2).P(B | A_2)}$$

RANDOM VARIABLE

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).

PROPERTIES OF A DISCRETE PROBABILITY DISTRIBUTION

- (1) $0 \leq P(X_i) \leq 1$ • for each X_i ($i = 1, 2, \dots, 7$)
- (2) $\sum p(X_i) = 1$

the mean is given by:

$$\mu = E(X) = \frac{\sum XP(X)}{\sum p(X)}$$

COMPUTATION OF THE STANDARD DEVIATION

Just as in case of a frequency distribution, we have

$$\begin{aligned} S &= \sqrt{\frac{\sum f(X - \bar{X})^2}{\sum f}} \\ &= \sqrt{\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2} = \sqrt{\frac{\sum X^2f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2} \end{aligned}$$

Similarly, in case of a probability distribution, we have

$$\begin{aligned} \sigma = \text{S.D.}(X) &= \sqrt{\frac{\sum X^2P(X)}{\sum P(X)} - \left[\frac{\sum XP(X)}{\sum P(X)}\right]^2} \\ &= \sqrt{\sum X^2P(X) - [\sum XP(X)]^2} \\ &\quad \left(\because \sum P(X) = 1\right) \end{aligned}$$

In the above example

Coefficient of Variation

$$C.V. = \frac{\sigma}{\mu} \times 100$$

DISTRIBUTION FUNCTION

Distribution function, mathematical expression that describes the probability that a system will take on a specific value or set of values. The classic examples are associated with games of chance.

$$F(x) = P(X \leq x)$$