INDEX

Pre-testing, data preparation, tabulation & examination

1. Meaning of Data processing

Data in the real world often comes with a large quantum and in a variety of formats that any meaningful interpretation of data cannot be achieved straightway. Social science researches, to be very specific, draw conclusions using both primary and secondary data.

To arrive at a meaningful interpretation on the research hypothesis, the researcher has to prepare his data for this purpose. This preparation involves the identification of data structures, the coding of data and the grouping of data for preliminary research interpretation. This data preparation for research analysis is teamed as processing of data.

Pre-testing, data preparation, tabulation & examination

Data processing is an intermediary stage of work between data collections and data interpretation. The data gathered in the form of questionnaires/interview schedules/field notes/data sheets is mostly in the form of a large volume of research variables. The research variables recognized, is the result of the preliminary research plan, which sets out the data processing methods beforehand.

Processing of data requires advanced planning and this planning may cover such aspects as identification of variables, hypothetical relationship among the variables and the

2. The various steps in processing of data

- ✓ Identifying the data structures
- ✓ Editing the data
- ✓ Coding and classifying the data
- ✓ Transcription of data
- ✓ Tabulation of data

Objectives

- Checking for analysis
- Editing
- Coding
- Classification
- Transcription of data
- Tabulation
- Construction of frequency table
- Measures of central tendency
- Components of a table
- Principles of table construction
- Frequency distribution & class intervals
- Graphs, charts and diagrams
- Types of graphs & general rules
- Quantitative & qualitative analysis
- Dispersion
- Correlation analysis

3. Checking for Analysis

In the data preparation step, the data are prepared in a data format, which allows the analyst to use modern analysis software such as SAS or SPSS. The major criterion in this is to define the data structure. **A data structure is a dynamic collection of related variables and can be conveniently represented as a graphic where nodes are labelled by variables.**

The data structure also defines & stages of the preliminary relationship between variables/groups that have been pre-planned by the researcher.

A sample structure could be a linear structure, in which one variable leads to the other and finally, to the resultant end variable.

4. Editing

The next step in the processing of data is editing of the data instruments. Editing is a process of checking to detect and correct errors and omissions. **Data editing happens at two stages, one at the time of recording of the data and second at the time of analysis of data.**

Data Editing at the Time of Recording of Data

All editing and cleaning steps are documented, so that, the re-definition of variables or later analytical modification requirements could be easily incorporated into the data sets.

5. Data editing at the time of analysis of data

Data editing is also a requisite before the analysis of data is carried out. This ensures that the data is complete in all respect for subjecting them to further analysis.

The editing steps checks for the completeness, accuracy, and uniformity of the data as created by the researcher.

(a) Completeness: The first step of editing is to check whether there is an answer to all the questions/variables set out in the data set. If there were any omission, the researcher sometimes would be able to deduce the correct answer from other related data on the same instrument.

Data editing at the time of analysis of data

(b) Accuracy: Apart from checking for omissions, the accuracy of each recorded answer should be checked. A random check process can be applied to trace the errors at this step. Consistency in response can also be checked at this step. The cross verification to a few related responses would help in checking for consistency in responses. The reliability of the data set would heavily depend on this step of error correction.

(c) Uniformity: In editing data sets, another keen lookout should be for any lack of uniformity, in interpretation of questions and instructions by the data recorders. For instance, the response towards a specific feeling could have been queried from a positive as well as a negative angle.

The final point in the editing of data set is to maintain a log of all collections that have been carried out at this stage. The documentation of these corrections helps the researcher to retain the original data set.

6. Coding

The edited data are then subject to codification & classification. Coding process assigns numerals or other symbols to the several responses of the data set. It is therefore a pre-requisite to prepare a coding scheme for the data set. The re-cording of the data is done on the basis of this coding scheme.

The responses collected in a data sheet varies, sometimes the responses could be the choice among a multiple response, sometimes the response could be in terms of values and sometimes in response could be alphanumeric. At the recording stage itself, if some codification were done to the responses collected, it would be useful in the data analysis. When codification is done, it is imperative to keep a log of codes allotted to the observations.

(a) Numeric Coding

Coding need not necessarily be numeric. It can also be alphabetic. Coding has to be compulsorily numeric, when the variables is subject to further parametric analysis.

(b) Alphabetic Coding: A mere tabulation or frequency count or graphical representation of the variable may be given in an alphabetic coding.

(c) Zero Coding: A coding of zero has to be assigned carefully to a variable. In many instances, when manual analysis done, a code of 0 would imply a 'no response' from the respondents.

The coding sheet needs to be prepared carefully, if the data recording is not done by the researcher, but is outsourced to a data entry firm or individual.

7. Classification

First, classification should be linked to the theory and the aim of the particular study. The objectives of the study will determine the dimensions chosen for coding.

Second, the scheme of classification should be exhaustive. That is, there must be category for every response.

Third, the categories must also be mutually exhaustive, so that each case is classified only once. This requirement is violated when some of the categories overlap or different dimensions are mixed up.

8. Transcription of Data

When the objectives collected by the researcher are not very large, the simple inferences, which can be drawn from the observations, can be transferred to a data sheet, which is a summary of all responses on all observations from a research instrument.

The transcription process helps in the presentation of all responses and observations on data sheets which can help the researcher to arrive at preliminary conclusions to the nature of the sample collected etc.

Transcription is an intermediary process between data coding & data tabulation.

9. Methods of Transcription

The researcher may adopt a manual or computerized transcription. Long work sheets, sorting cards or sorting strips could be used by the researcher to manually transcript the responses. The computerized transcription could be done using a data base package such as spreadsheets, text files or other databases.

Manual Transcription: When the sample size is manageable, the researcher need not use any computerization process to analyze the data. The researcher could prefer a manual transcription and analysis responses. The choice of manual transcription would be when the number of responses in a researcher instrument is very less.

10. Long Worksheets

Long worksheets require quality paper, preferably chart sheets, thick enough to last several usages. These worksheets normally are ruled both horizontally and vertically, allowing responses to be written in the boxes. If one sheet is not sufficient, the researcher may use multiple rules sheets to accommodate all the observations.

Tabulation: The transcription of data can be used to summarize and arrange the data incompact form for further analysis. The process is called tabulation. Tabulation is a process of summarizing raw data displaying them on compact statistical tables for further analysis. It involves counting the number of cases falling into each of the categories identified by the researcher.

11. Manual Tabulation

When data are transcribed in a classified form as per the planned scheme of classification, category-wise totals can be extracted from the respective columns of the work sheets.

Computerized tabulation is easy with the help of software packages. The input requirement will be the column and row variables. The software package then computes the number of records in each cell of three row column categories. The most popular package is the Statistical package for Social Science (SPSS).

12. Construction of Frequency Table

Frequency tables provide a "shorthand" summary of data. The importance of presenting statistical data in tabular form needs no emphasis.

The general purpose tables are primary or reference tables designed to include large amount of source of data in convenient and accessible form. The special purpose tables are analytical or derivate ones accessible form. The special purpose tables are analytical or derivative ones that demonstrate significant relationships in the data or the results of statistical analysis. In research, we are primarily concerned with special purpose.

13. Components of a Table

The major components of a table are

A. Heading

(a) Table Number

(b) Title of the Table

(c) Designation of units

B. Body

i. **Sub-head:** Heading of all rows or blocks of stub items

ii. **Body-head:** Headings of all columns or main captions and their sub-captions

iii. **Field body:** The cells in rows and columns

Components of a Table

C. Notations

I. Footnotes, wherever applicable.

II. Source, wherever applicable.

14. Frequency distribution & class intervals

Variables that are classified according to magnitude or size are often arranged in the form of a frequency table. In constructing this table, it is necessary to determine the number of class intervals to be used and the size of the class intervals.

A distinction is usually made between continuous and discrete variables. A continuous variable has an unlimited number of possible values between the lowest and highest with no gaps or breaks.

Frequency distribution and class intervals

Class Intervals: Ordinarily, the number of class intervals may not be less than 5 not more than 15, depending on the nature of the data and the number of cases being studied. After noting the highest and lower values and the feature of the data, the number of intervals can be easily determined.

One-way Table: One-way frequency tables present the distribution of cases on only a single dimension or variable. For instance, the gender distribution of a sample study may be described as "The sample data represents 58% by males and 42% of the sample are females."

Frequency distribution and class intervals

Category Members	Extent of participation						
	Low No. of Respondents	%	Medium No. of Respondents	%	High No. of Respondents	%	Total
Ordinary	65	41.9	83	56.8	2	1.3	115
Committee	4	10.3	33	84.6	2	5.1	39

Two-Way Table: Distribution in terms of two or more variables and then relationship between the two variables are show in two-way table. The categories of one variable are presented one below another, on the left margin of the table those of another variable at the upper part of the table, one by the side of another. The cells represent particular combination of both variables. To compare the distributions of cases, raw numbers are converted into percentage based on the number of cases in each category.

Frequency distribution and Class intervals

Economic Status	Democratic participation			
	Low	Medium	High	Total
Low	6 (35.3)	11 (64.7)	0 (0.0)	17
Medium	13 (38.2)	18 (53.0)	3 (8.8)	34
High	6 (62.5)	10 (62.5)	0 (0.0)	16
Very High	2 (33.3)	3 (50.0)	1 (16.7)	6
Total	27	42	4	73

Another method of constructing a two-way table is to state the percent of representation as a within brackets term rather than as a separate column. Here, special care has been taken as to how the percentages are calculated, either on a horizontal representation of data or as vertical representation of data.

15. Graphs, Charts & Diagrams

In presenting the data of frequency distributions and statistical computations, it is often desirable to use appropriate forms of graphics, charts and other pictorial devices such as diagrams.

The meaning of figures in tabular form may be difficult for the mind to grasp or retain. "Properly constructed graphs and charts relieve the mind of burdensome details by portraying facts concisely, logically and simply."

Graphic presentation must be planned with utmost care and diligence. Graphic forms used should be simple, clear and accurate and also appropriate to the data.

16. Types of Graphs and general rules

a) Line graphs or charts
b) Bar charts
c) Segmental presentations
d) Scatter plots
e) Bubble charts
f) Stock plots
g) Pictographs
h) Chesnokov faces

The general rules to be followed in graphic representations are

a. The chart should have a title placed directly above the chart
b. The title should be clear, concise and simple and should be presented in an accompanying table
c. Numerical data upon which the chart is based should be presented in an accompanying table

Types of Graphs and General Rules

d. The horizontal line measures time or independent variable and the vertical line the measured variable

e. Measurements proceed from left to right on the horizontal line and from bottom to top on the vertical

f. Each curve or bar on the chart should be labeled

g. If there are more than one curves or bar, they should be clearly differentiated from one another by distinct patterns or colors

h. The zero point should always be represented and the scale intervals should be equal

i. Graphic forms should be used sparingly. Too many forms detract rather than illuminating the presentation

j. Graphic forms should follow and not precede the related textual discussion

17. Line Graphs

The line graph is useful for showing changes in data relationship over a period of time. In this graph, figures are plotted in relation to two intersecting lines or axes. The horizontal line is called the abscissa or X-axis and the vertical, the ordinal or Y-axis. The point at which the two axes intersect is zero for both X and Y axis. The 'O' is the origin of coordinates.

The two lines divide the region of the plane into four sections known as quadrants that are numbered anti-clockwise. Measurements to the right and above "O" are positive and measurements to the left and below "O" are negative, is an illustration of the features of a rectangular coordinate type of graph. Any point of plane of the two axes is plotted in terms of the two axes reading from the origin "O". Scale intervals in both the axes should be equal.

Line Graphs

If a part of the scale is omitted, a set of parallel jagged lines should be used to indicate the break in the scale. The time dimension or independent variable is represented by the X-axis and the other variable by Y-axis.

TABLE – 1 ASEAN- Demographic Factors

	Land Area (Sq.Km)	Popula-tion (000)	Popula-tion density (persons per sq km)	Sex ratio (males per 100 females)	Popula-tion below 5 years (000)	Popula-tion 65 years and over (000)	Popula-tion 15-29 years (000)	Urban popula-tion (%)	Popula-tion Living below national poverty line (%)	Popula-tion living below PPP $1.9 (%)	Unempl-oyment rate (%),	Adult literacy rate (%)	Infant mortality rate (per 1000 live births)	Life expec-tancy (years)
Brunei-Darussalam	5,765	442	76.7	111.7	30.1	8.1	121.9	77.6	NA	NA	9.2	97.2	9	77.5
Cambodia	181,035	15,982	88.3	95.2	1,770.9	340	4,687.3	23.4	13.5	24	1.1	82.5	25.1	70.6
Indonesia	1,916,862.2	265,015	138.3	91.8	23,920.3	7,836.5	63,254	55.3	9.8	5.7	5.3	95.5	21.4	71.2
Lao PDR	236,800	6,887	29.1	100.6	723.5	117	2,088.8	35	23.4	22.7	0.6	84.7	48.6	67
Malaysia	331,388	32,385	97.7	106.8	2,595.6	783.3	9,384.6	75.6	0.4	0	3.3	95.1	6.7	75
Mayanmar	676,576	53,625	79.3	92.4	4,986	1316.6	13,743	30.6	24.8	6.4	1	89.6	38.5	66.7
Philippines	300,000	106,599	355.3	101.8	11,431.7	2080.2	29,134.1	46.9	21.6	8.3	5.4	96.4	22.2	69.2
Singapore	720	5,639	7,832.6	95.9	185.5	205.6	773.7	100	NA	NA	2.9	97.2	2.2	83.2
Thailand	513,140	67,832	132.2	95	3,743.3	2730.5	13,803.1	49.9	7.9	0	1.1	96.1	8.2	75.5
Vietnam	331,230	94,666	285.8	97.7	7,263.3	3367.7	19,432.6	35.7	9.8	2	2.2	95.1	16.7	73.5

18. Measures of Location

The farthest one can reduce a set of data, and still retain any information at all, is to summarize the data with a single value. Measures of location do just that: They try to capture with a single number what is typical of the data. What single number is most representative of an entire list of numbers? We cannot say without defining "representative" more precisely. We will study three common measures of location: the mean, the median, and the mode. The mean, median and mode are all "most representative," but for different, related notions of representatives.

For qualitative and categorical data, the mode makes sense, but the mean and median do not.

It is hard to see the connection between the mean, median, and mode from their definitions.

Measures of Location

However, the mean, the median, and the mode are "as close as possible" to all the data: For each of these three measures of location, the sum of the distances between each datum and the measure of location is as small as it can be. The differences among the three measures of location are in how "distance" is defined.

- For the mean, the distance between two numbers is defined to be the square of their difference.
- For the median, the distance between two numbers is defined to be the absolute value of their difference.
- For the mode, the distance between two numbers is defined to be zero if the numbers are equal, & one if they are not equal.

Measures of Location

The mean, median, and mode can be related (approximately) to the histogram: loosely speaking, the mode is the highest bump, the median is where half the area is to the right and half is to the left, and the mean is where the histogram would balance, were it a solid object cut out of a uniform block of metal.

Central Tendencies – Central tendency is a measure that characterizes the central value of a collection of data that tends to cluster somewhere between the high and low values in the data. It refers to measurements like mean, median and mode. It is also called measures of center.

19. Mean

The mean is the most common measure of central tendency. It is the ratio of the sum of the scores to the number of the scores. For ungrouped data which has not been grouped in intervals, the arithmetic mean is the sum of all the values in that population, divided by the number of values in the population as where, μ is the arithmetic mean of the population, Xi is the ith value observed, N is the number of items in the observed population and Σ is the sum of the values. For example, the production of an item for 5 days is 500, 750, 600, 450 and 775 then the arithmetic mean is μ = 500 + 750 + 600 + 450 + 775/ 5 = 615.

Mean

Weighted Mean – When a mean is calculated, a serious mistake can be committed if one overlooks the fact that the quantities that are being averaged are not all of equal importance with reference to the situation being described.

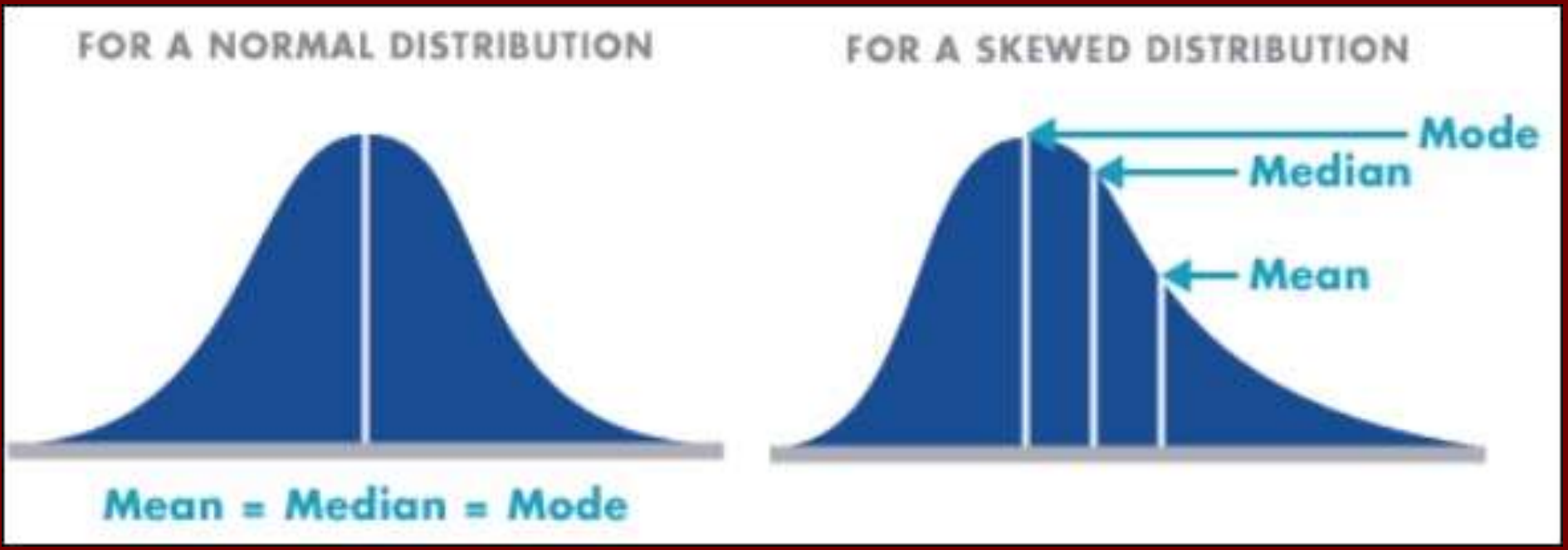
Median

It divides the distribution into halves; half are above it and half are below it when the data are arranged in numerical order. It is also called as the score at the 50th percentile in the distribution. The median location of N numbers can be found by the formula (N + 1) / 2. When N is an odd number, the formula yields an integer that represents the value in a numerically ordered distribution corresponding to the median location.

20. Mode

It is the most frequent or common score in the distribution or the point or value of X that corresponds to the highest point on the distribution. If the highest frequency is shared by more than one value, the distribution is said to be multimodal and with two, it is bimodal or peaks in scoring at two different points in the distribution.

In general, the mean and the median need not be close together. If the data have a symmetric distribution, the mean and median are exactly equal, but if the distribution of the data is skewed, the difference between mean and the median can be large. The median is smaller than the mean if the data are skewed to the right, and larger than the mean if the data are skewed to the left.


FOR A NORMAL DISTRIBUTION
FOR A SKEWED DISTRIBUTION
Mode
Median
Mean
Mean = Median = Mode

21. Measures of Shape: Skewness and Kurtosis

The measure of central tendency and measure of dispersion can describe the distribution but they are not sufficient to describe the nature of the distribution. For this purpose, we use other two statistical measures that compare the shape to the normal curve called Skewness and Kurtosis.

Skewness and Kurtosis are the two important characteristics of distribution that are studied in descriptive statistics.

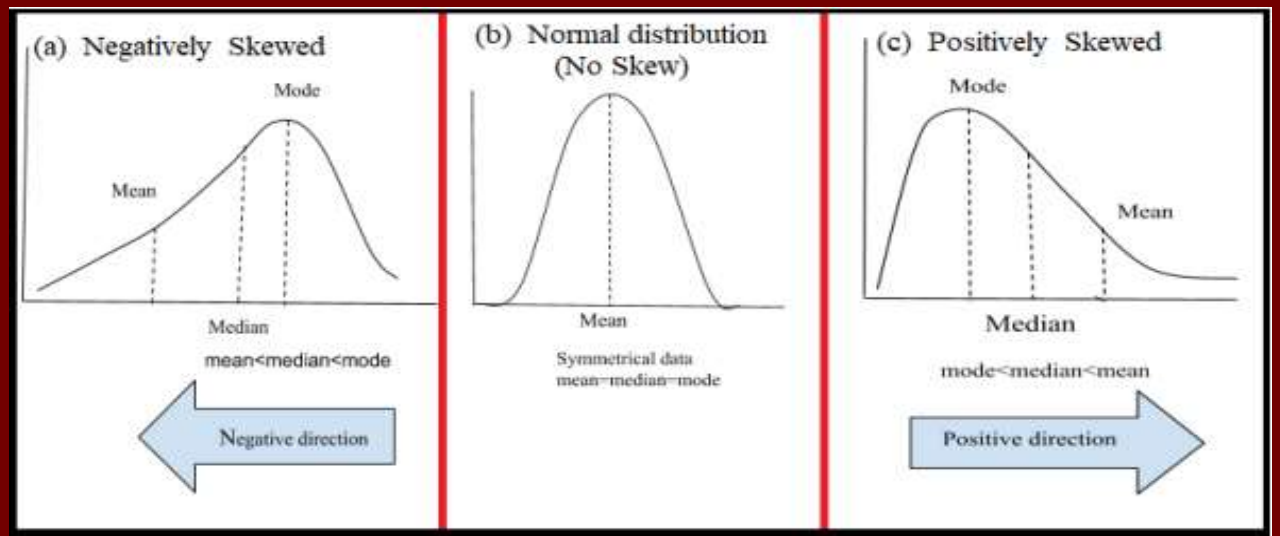
22. Skewness

Skewness is a statistical number that tells us if a distribution is symmetric or not. A distribution is symmetric if the right side of the distribution is similar to the left side of the distribution.

Skewness

If a distribution is symmetric, then the Skewness value is 0. i.e. If a distribution is Symmetric (normal distribution): median= mean= mode, (Skewness value is 0) If Skewness is greater than 0, then it is called right-skewed or that the right tail is longer than the left tail. If Skewness is less than 0, then it is called left-skewed or that the left tail is longer than the right tail.

For example, the symmetrical and skewed distributions are shown by curves as



Skewness

The Formula of Skewness is

$$\text{Skewness} = \frac{\sum \left(x - \bar{x}\right)^3}{(n-1) \cdot S^3}$$

Where:
S: standard deviation
$\bar{X}$: Mean

Kurtosis

Kurtosis is a statistical number that tells us if a distribution is taller or shorter than a normal distribution. If a distribution is similar to the normal distribution, the Kurtosis value is 0. If Kurtosis is greater than 0, then it has a higher peak compared to the normal distribution. If Kurtosis is less than 0, then it is flatter than a normal distribution.

23. Kurtosis

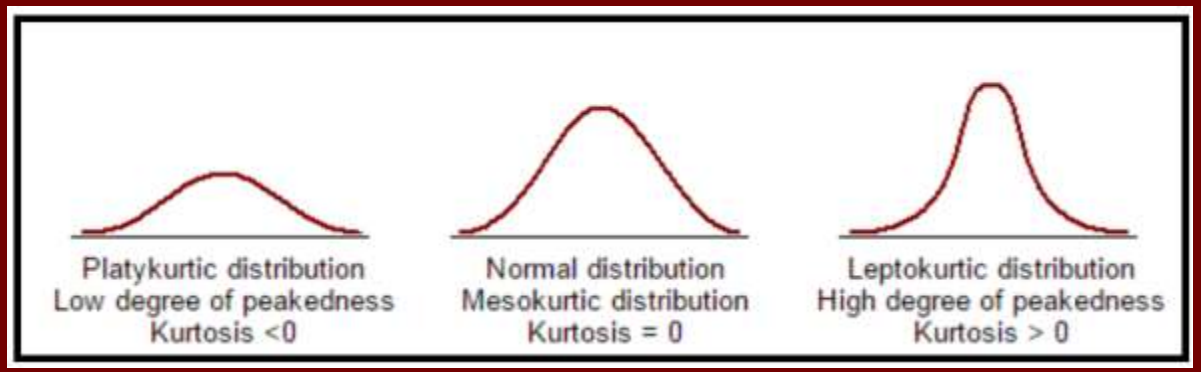
There are three types of distributions

Leptokurtic: Sharply peaked with fat tails, and less variable

Mesokurtic: Medium peaked

Platykurtic: Flattest peak and highly dispersed

For example, The different types of Kurtosis



Kurtosis

The Formula of kurtosis is:

$$\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n - 1) \cdot S^4}$$

Where:

S: standard deviation $\bar{X}$: Mean

24. Key differences between Skewness and Kurtosis

This is the fundamental differences between skewness and kurtosis

A- The characteristic of a frequency distribution that ascertains its symmetry about the mean is called skewness. Kurtosis means the relative pointedness of the standard bell curve, defined by the frequency distribution.

B- Skewness is a measure of the degree of lopsidedness in the frequency distribution. Conversely, kurtosis is a measure of degree of tailed-ness in the frequency distribution.

Key differences between Skewness and Kurtosis

C- Skewness is an indicator of lack of symmetry, i.e. both left and right sides of the curve are unequal, with respect to the central point. As against this, kurtosis is a measure of data, that is either peaked or flat, with respect to the probability distribution.

D- Skewness shows how much and in which direction, the values deviate from the mean? In contrast, kurtosis explain how tall and sharp the central peak is.

25. Measures of Variability

There are many ways to describe variability or spread including

- **Range**
- Interquartile range (IQR)
- **Variance and Standard Deviation**

Range: The range is the difference in the maximum and minimum values of a data set. The maximum is the largest value in the dataset and the minimum is the smallest value. The range is easy to calculate but it is very much affected by extreme values.

$$Range = maximum - minimum$$

Measures of Variability

Interquartile Range (IQR): The interquartile range is the difference between upper and lower quartiles and denoted as IQR.

$$\begin{aligned} IQR &= Q3 - Q1 \\ &= upper\ quartile - lower\ quartile \\ &= 75th\ percentile - 25th\ percentile \end{aligned}$$

Variance and Standard Deviation Section

One way to describe spread or variability is to compute the standard deviation. The standard deviation is the square root of the variance.

Measures of Variability

Variance

the average squared distance from the mean

Population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

where μ is the population mean and the summation is over all possible values of the population and N is the population size.

σ^2 is often estimated by using the sample variance.

Sample Variance

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

Where n is the sample size and $\bar{x}$ is the sample mean.

26. Standard Deviation

The standard deviation is a very useful measure. One reason is that it has the same unit of measurement as the data itself (e.g. if a sample of student heights were in inches then so, too, would be the standard deviation.

Approximately the average distance the values of a data set are from the mean or the square root of the variance.

Population Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

It has the same unit as the x_i's. This is a desirable property since one may think about the spread in terms of the original unit.

σ is estimated by the sample standard deviation s :

Sample Standard Deviation

$$s = \sqrt{s^2}$$

A rough estimate of the standard deviation can be found using $s \approx \frac{\text{range}}{4}$

Coefficient of Variation

A popular statistic to use in such situations is the Coefficient of Variation or CV. This is a unit-free statistic and one where the higher the value the greater the dispersion. The calculation of CV is

Coefficient of Variation (CV)

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}}$$