

PROBABILITY AND STATISTICS


Prepared by:
Sudip Pokhrel

Several thin, parallel white lines of varying lengths and orientations are located in the bottom right corner of the slide, creating a modern, abstract design element.

DEFINITION OF STATISTICS

- ▶ Statistics are branch of science where we plan, gather, and analyze information about a particular collection of individuals or objects under investigation.
- ▶ Statistics is the study of methods and procedures for collection, classification, analysis and interpretation of quantitative data to make scientific inference.
- ▶ Prof. Horace Secrist states that '*Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a symmetric manner, for a predetermined purpose and placed in relation to each other.*'

Characteristics of Statistics

- ▶ Statistics are always expressed numerically.
 - ▶ Statistics are the aggregate of facts.
 - ▶ Statistics are influenced by multiplicity of causes.
 - ▶ It should be done in symmetric manner.
 - ▶ Statistical data is collected for pre determined purpose.
 - ▶ In statistics, there is certain degree of accuracy.
- 

Major two Components of Statistics

► Descriptive Statistics

- Exploratory Data Analysis
- Gives numerical and graphical procedures to summarize a collection of data in a clear and understandable way.

Goal: visualize relationships, generate hypotheses

► Inferential Statistics

- Confirmatory Data Analysis
- Hypothesis tests
- Confidence Intervals
- Regression modeling
- Provides procedures to draw inferences about a population from a sample.

Goal: quantify relationships, test hypotheses



Key Descriptive Statistics Ideas

Descriptive statistics helps to simplify large amounts of data in a sensible way. Each descriptive statistics reduces lots of data into a simpler summary

□ **Graphical Method(Visualizing data)**

- Bar Diagrams
- Histograms
- Boxplots
- Scatterplots
- Pie chart etc.

□ **Numerical Method(Describing data)**

- Distribution shapes (especially skewness)
- Quartiles
- Measures of central tendency

Median, Mean, Mode

- Measures of spread


Variance, Standard Déviation, Range, Interquartile Range(IR)

Descriptive Statistics Vs Inferential Statistics


BASIS FOR COMPARISON	DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Meaning	Branch of statistics which is concerned with describing the population under study.	A type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Conclusion	<ul style="list-style-type: none">- It is all about illustrating your current dataset whereas inferential statistics focuses- It provides summation of the data the researcher has actually studied	<ul style="list-style-type: none">- Inferential statistics focuses on making assumptions on the additional population, that is beyond the dataset under study.- Inferential statistics, makes the generalization, which means the data provided to you is not actually studied.

FUNCTIONS OF STATISTICS:

- ▶ To represent facts from numerical figures in a finite form.
 - ▶ It simplifies complexity.
 - ▶ To help classification of data.
 - ▶ To provide methods for making comparison.
 - ▶ To help formulating policies.
- 
- A series of several parallel white diagonal lines of varying lengths, located in the bottom right corner of the slide, extending from the bottom edge towards the right edge.

- ▶ To determining relationship between different phenomena.
 - ▶ To help predicting future trends.
 - ▶ To formulate and test the hypothesis.
 - ▶ To have an idea about the occurrence or non occurrence of certain events.
 - ▶ To draw valid inferences or conclusions.
- 

LIMITATIONS OF STATISTICS

- ▶ Statistics does not deal with individual.
 - ▶ Statistical technique deals with the quantitative data only. It ignores qualitative aspects like beauty, goodness, knowledge etc.
 - ▶ Statistical laws are not exact.
 - ▶ Statistics can be misused.
- 
- A series of four parallel white diagonal lines in the bottom right corner of the slide, slanting upwards from left to right.

SIGNIFICANCE OF STATISTICS IN ENGINEERING

► **Design of experiment:**

Statistics are used for deriving techniques to model the experiment for component construction and system

▣ **Quality and process control**

Uses statistics as a tool to maintain conformance of the system to its intended functionalities for the manufacturing process and it has also control over the process of manufacture

► **Time and method engineering**

Uses statistics for the study of repeated operation in the manufacturing process to maintain standard and also for optimal (in some cases)

► **Reliability Engineering**

Used in reliability engineering to measure the ability of machines/equipment's to perform as per its specifications and it also measure for improvement.

► **Probabilistic Design**

Used for probabilistic design for the design of components and system in any process

What is Data(singular 'Datum')?

- Data are a collection of observations about the variable being measured.
- Data means the actual observation of the phenomenon of interest based on sample from the population.
- It is either Quantitative(Numerical) or Qualitative(Categorical).

Eg: -The pH level of yogurt. - Number of children per family. -A asthma patient is either smoker or non-smoker. - Type of electrical devices

What is Variable ?

- A variable is a characteristic that changes(i.e. show variability) from unit to unit(subject, plot, etc..)
- Data are the observation of Variables.

Scale of Measurement

► ***Data on a Nominal Scale:***

It consists of “naming” observation or classifying them into various mutually exclusive and collectively exhaustive categories.

E.g. : - Occupation can be classified as Student, Doctor, Engineer, Lecturer etc.

► ***Data on an Ordinal Scale***

Observations are not only different from the categories but can be ranked according to some criterion . E.g. :

► ***Data on an Interval Scale***

Data can be ranked, but now distances between two adjacent points on scale will be the same. E.g. :

► ***Data on an Ratio Scale***

A ratio scale is an interval scale with a true zero point. There is meaning to saying that an observation is x times larger, longer, heavier, faster, etc. than another. E.g. :

Collection of data

1. Primary data

(questionnaire survey, interview, Computer assisted personal interview, mailed questionnaire, focus group discussion , observation , case studies, dairies etc.)

2. secondary data

i. Some journals, national level, public sources

Like Nepal census of household and population, agriculture, business data, vital statistics, UN publication, budgets etc.

ii. Unusual, but easily accessible community data sources
like topographic maps, newspaper, films, post cards etc.

Methods of data collection

1. Census

2. Sample method

Diagrammatic and graphical Representation of Data


A diagram is a two dimensional geometric (can be three dimensional also) symbolic representation of information according to some visualization technique

A graph is a visual representation of a relationship between, but restricted to, two variables. A graph generally takes the form of one-dimensional or two dimensional figure. A graph commonly consists of two axes called the horizontal and vertical (X and Y- axis). Each axis corresponds to one variable. Each point on graph is defined by pair of numbers containing two coordinates (x and y). The point on graph represents relationship between the given variables

Importance of Diagram and Graph

- It give a bird's eye view of a set of numerical data and present in simple form
- A clear picture of variation in the values of a variable is much more easily obtained by diagrams and graphs than the value given in the table
- They can easily understood even by a common person
- They give delight to eye and leave an ever lasting impression on the mind
- They give required information in less time and without any mental strain
- They are useful to compare the two or more sets of numerical data

Limitations of Diagrams and graphs

- There is a loss of accuracy of data while representing data through diagrams. It is obvious that there will be loss of data as it is summarization of whole data. Due to this, the comparison between the values of data are not so accurate
 - Sometimes the illusionary effect creates a wrong impression on the mind of viewer
 - For some of them, constructing graphs may be time consuming or costly
- 

Types of diagrams

- **One dimensional**

- Simple bar diagram, Collateral or multiple bar diagram, sub divided or component bar diagram, percentage diagram

- **Two dimensional**

- Rectangles, squares, pie- diagrams

- **Three dimensional**

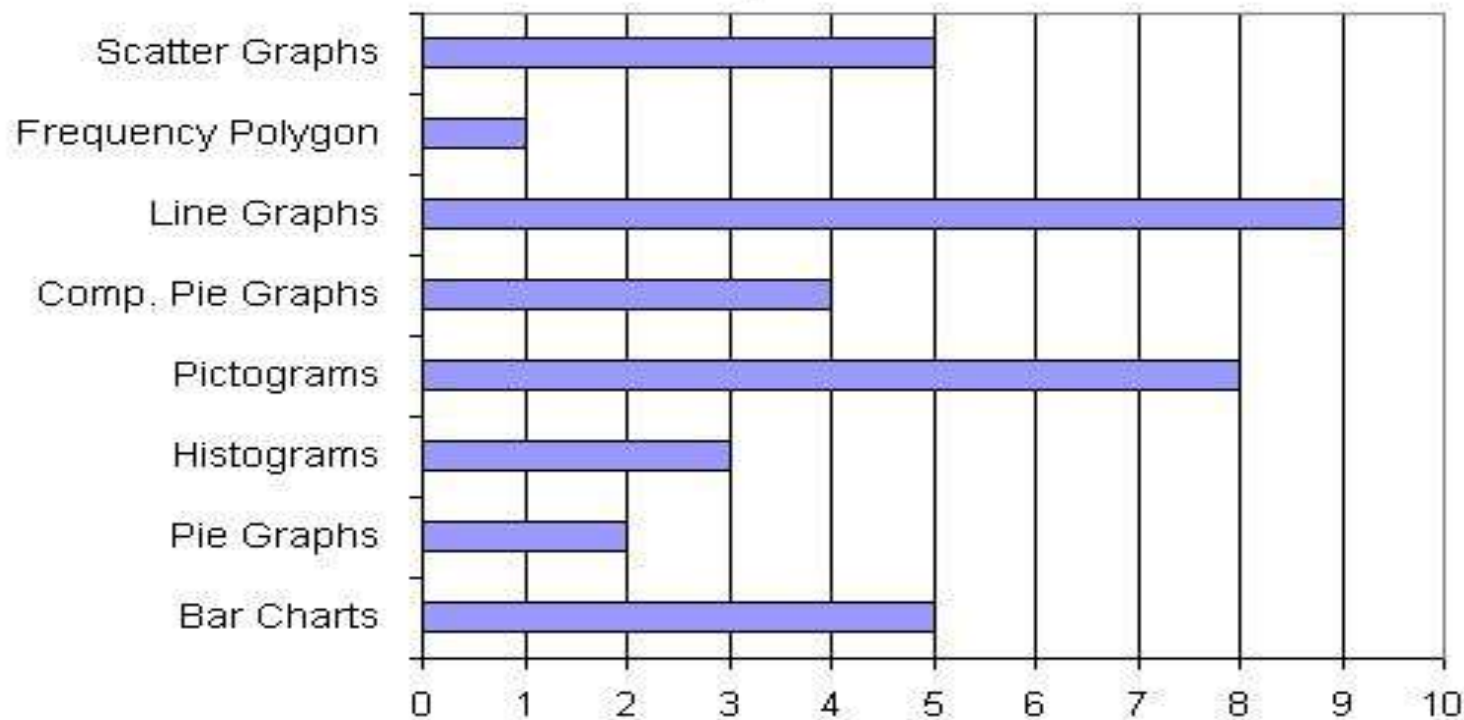
Known as volume diagrams consists of cubes, cylinders, spheres etc. (length, width and height have to taken into account while constructing diagrams)

Graphic representation of data

- **Graph of time series:** graphic representation of chronological data (varying according to time)
- **Graph of frequency Distribution:** Histogram, Frequency polygon, frequency curve, Ogive or cumulative frequency curve

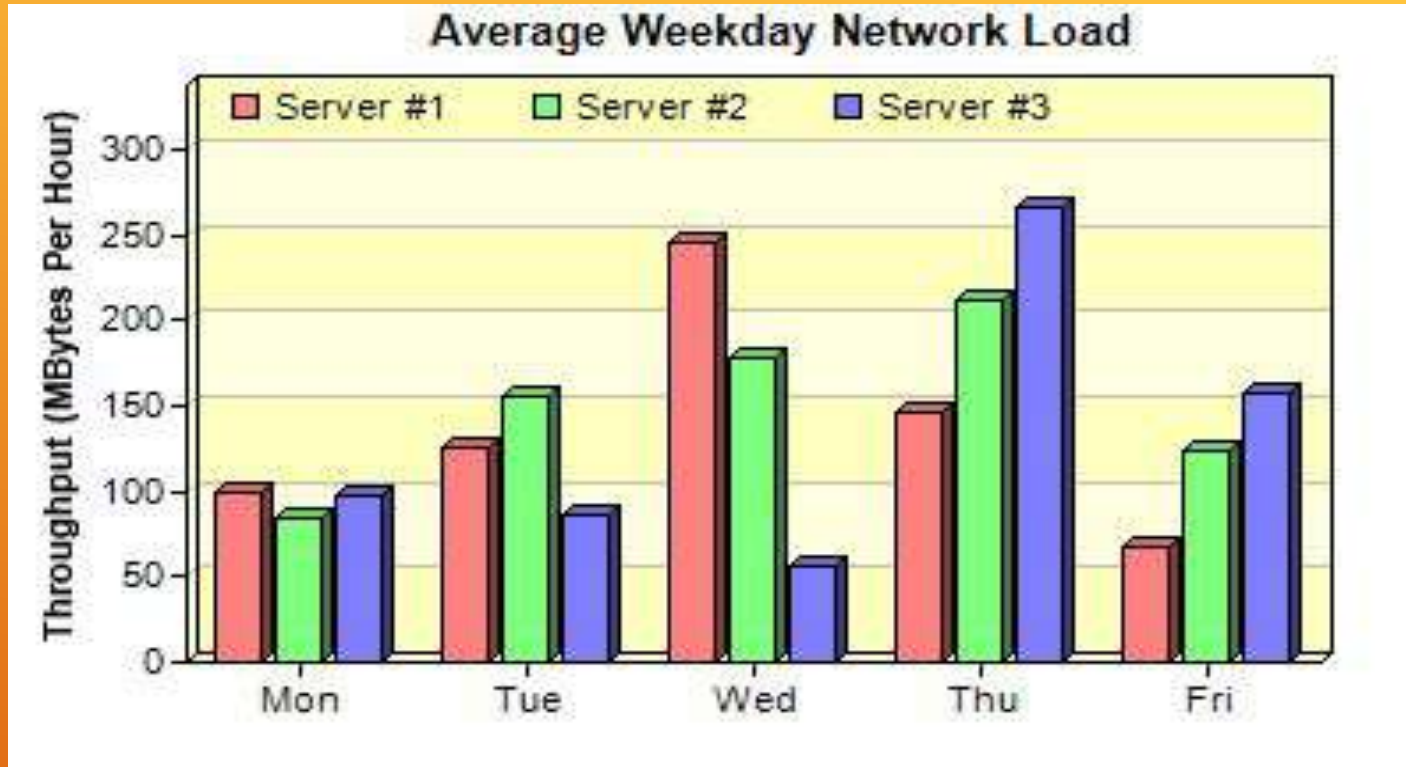
Bar Diagram

Favorite Graphs



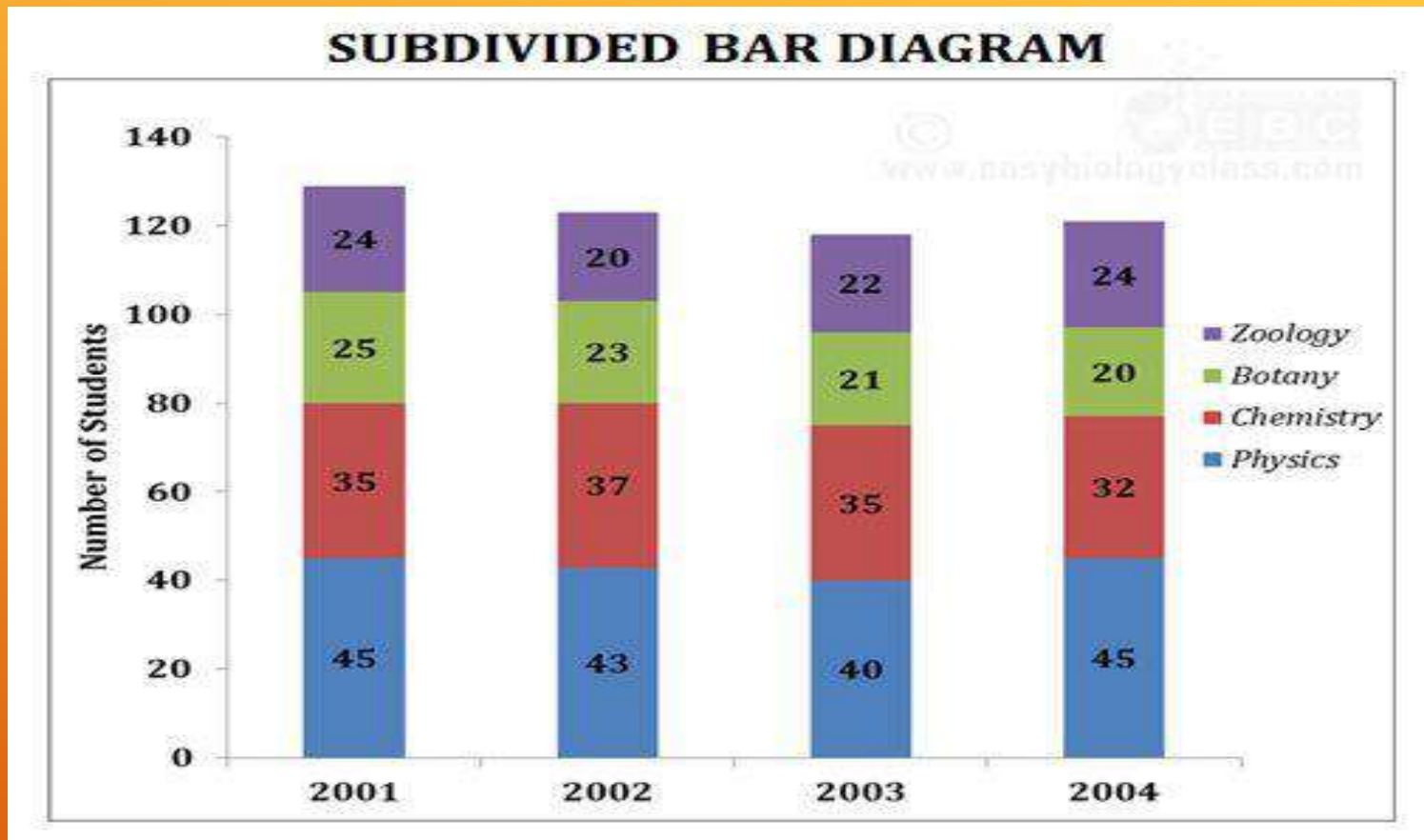
- Consists of set of rectangles of equal width
- The heights (or length) of the rectangles represented by the given values of variables
- The number of rectangle is equal to the number of the single variable
- Used for comparative study of two or more values of a single variable

Multiple Bar Diagram



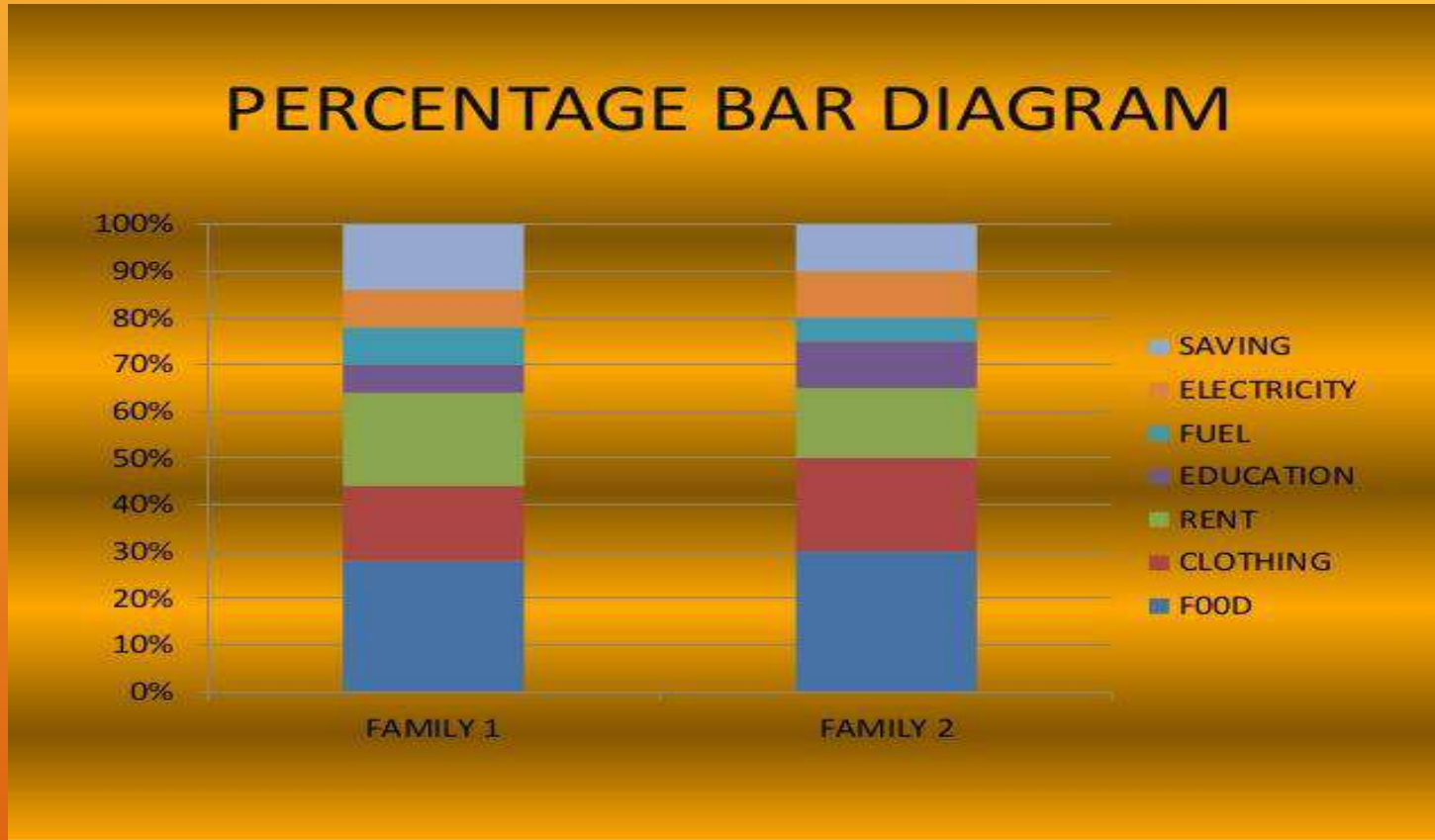
- Used to comparing for two or more sets of statistical data
- Bars are constructed side by side to represent the set of values for comparison
- Each bar in a group is shaded or coloured differently for the sake of distinction.

Subdivided or Component Bar Diagram



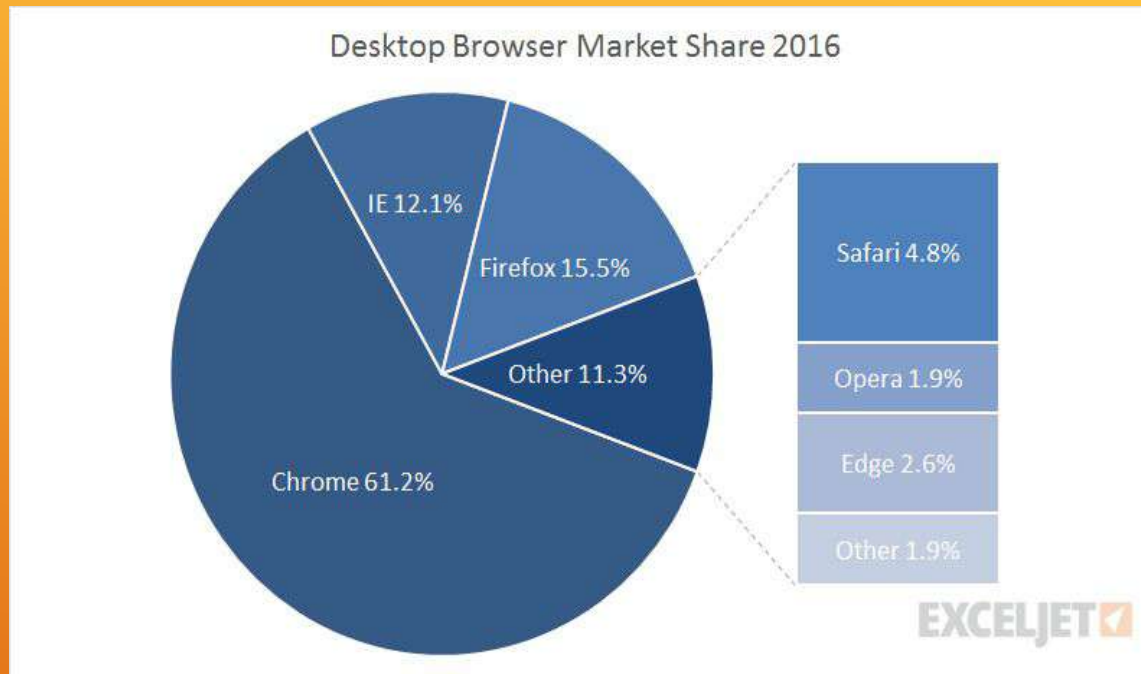
- This chart consists of bars which are sub-divided into two or more parts.
- The length of the bars is proportional to the totals.
- The component bars are shaded or colored differently.

Percentage Bar Diagram



- Component bar charts may also be drawn on percentage basis by expressing the components as percentages of their respective totals.
- All the bars are of equal length showing the 100%. These bars are sub-divided into component bars in proportion to the percentages of their components

Pie Chart or Circular Diagram

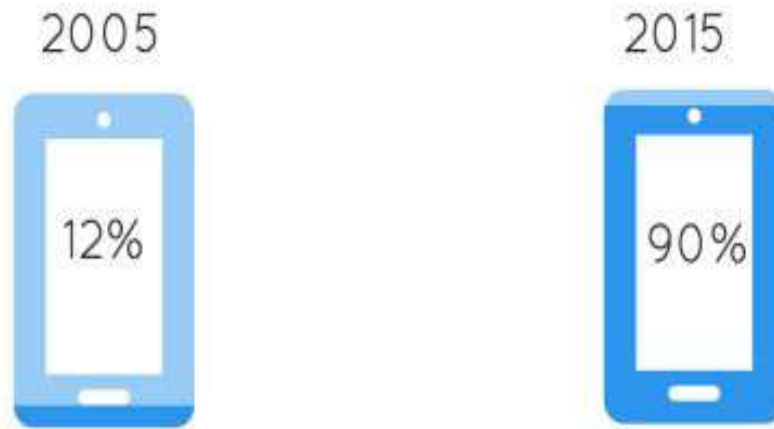


1. Pie chart is used to compare the relation between the whole and its components.
2. The difference between the component bar chart and pie chart is that in case of component bar chart the length of the bars are used while in case of a pie chart the area of the sector of a circle is used.
3. In pie chart, the circle is drawn with radii proportional to the square root of the quantities to be represented because the area of a circle is given by $2\pi r^2$.
4. The sectors are colored and shaded differently.
5. To construct a pie chart, we draw a circle with some suitable radius (square root of the total). The angles are calculated for each sector as follows:

$$\text{Angles for each sector} = \frac{\text{Component Part}}{\text{Total}} \times 360^\circ$$

Pictogram

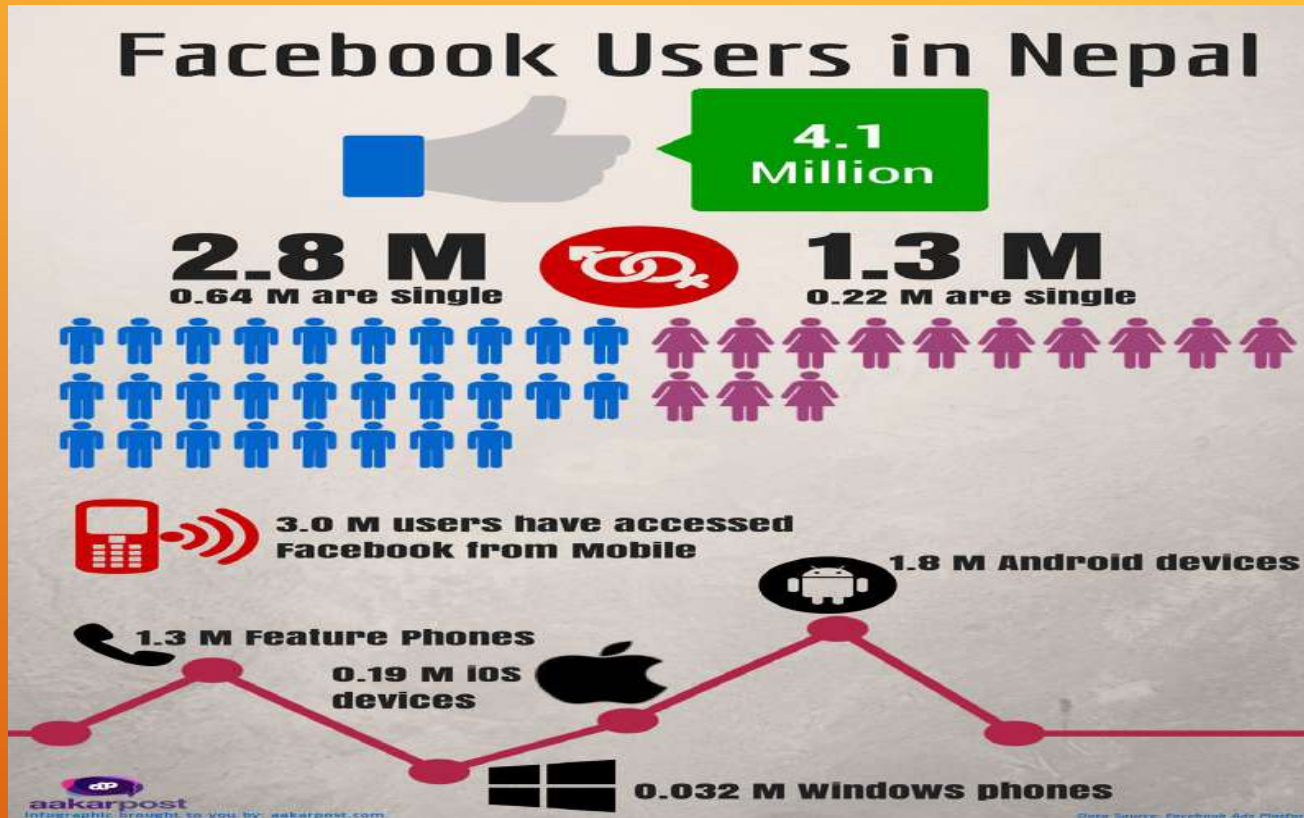
Percentage of Adults Aged 18-29 Who Use Social Media



Source: <http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/>

- Pictures are attractive and easy to comprehend and this method is particularly useful in presenting statistics to the layman.
- Data presented through a pictorial symbol that is carefully selected

Infographics

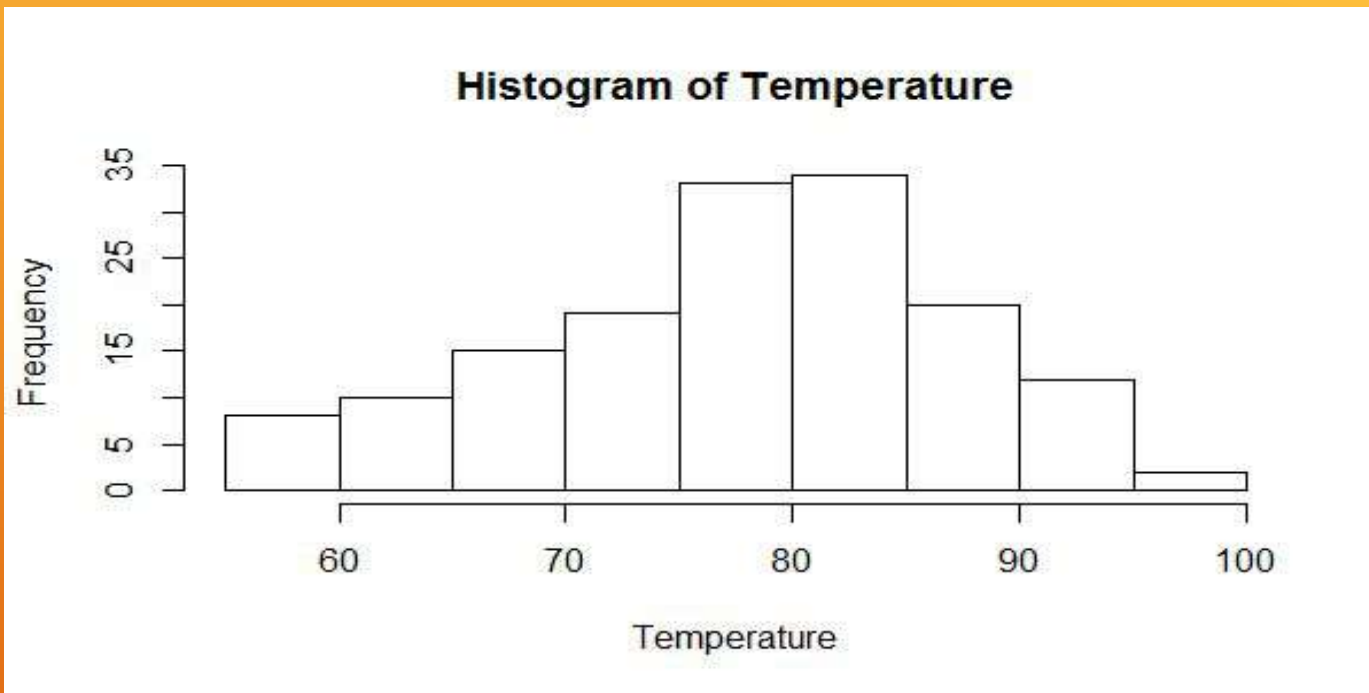


Infographics (a clipped compound of "information" and "graphics") are graphic visual representations of information, data or knowledge intended to present information quickly and clearly.

They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends.

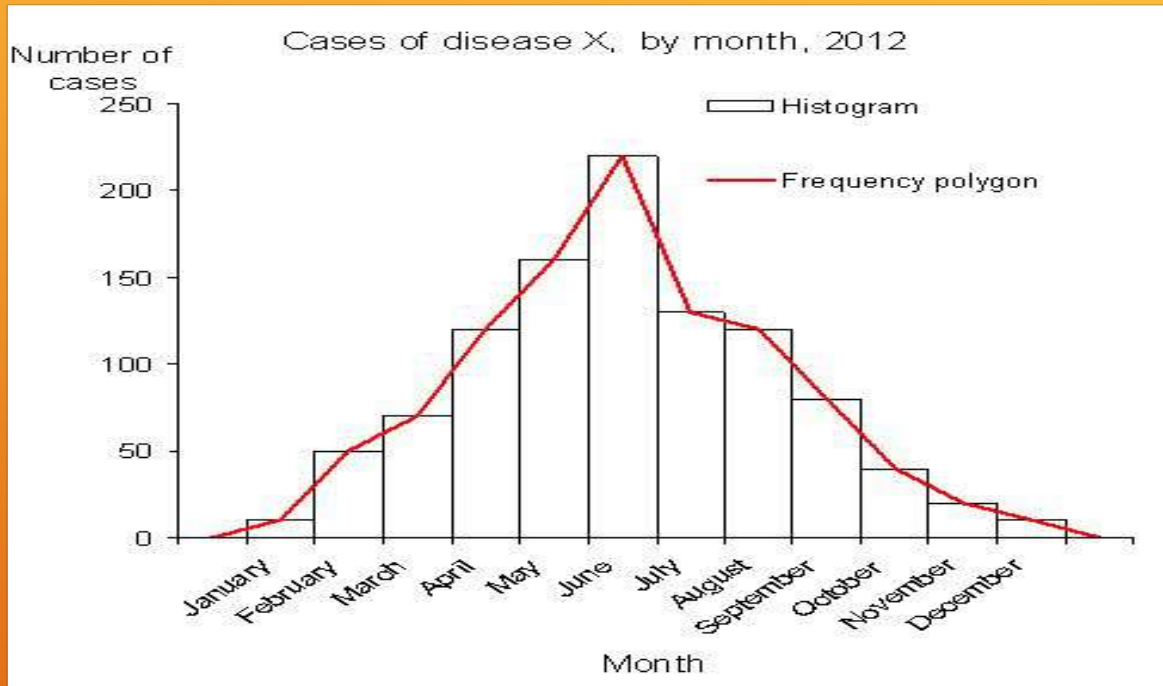
Similar pursuits are information visualization, data visualization, statistical graphics, information design, or information architecture.

Histogram



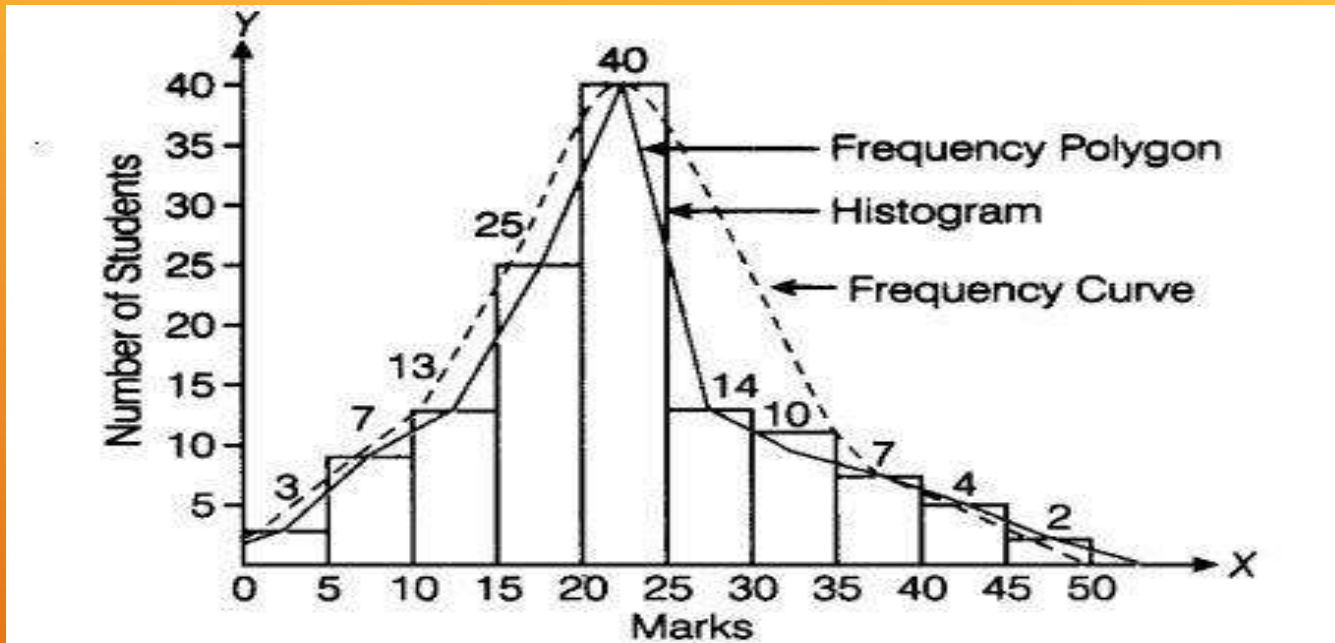
- A histogram displays the distribution of the data visually by representing the frequency or likelihood of the values in the sample.
- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes

Frequency Polygon



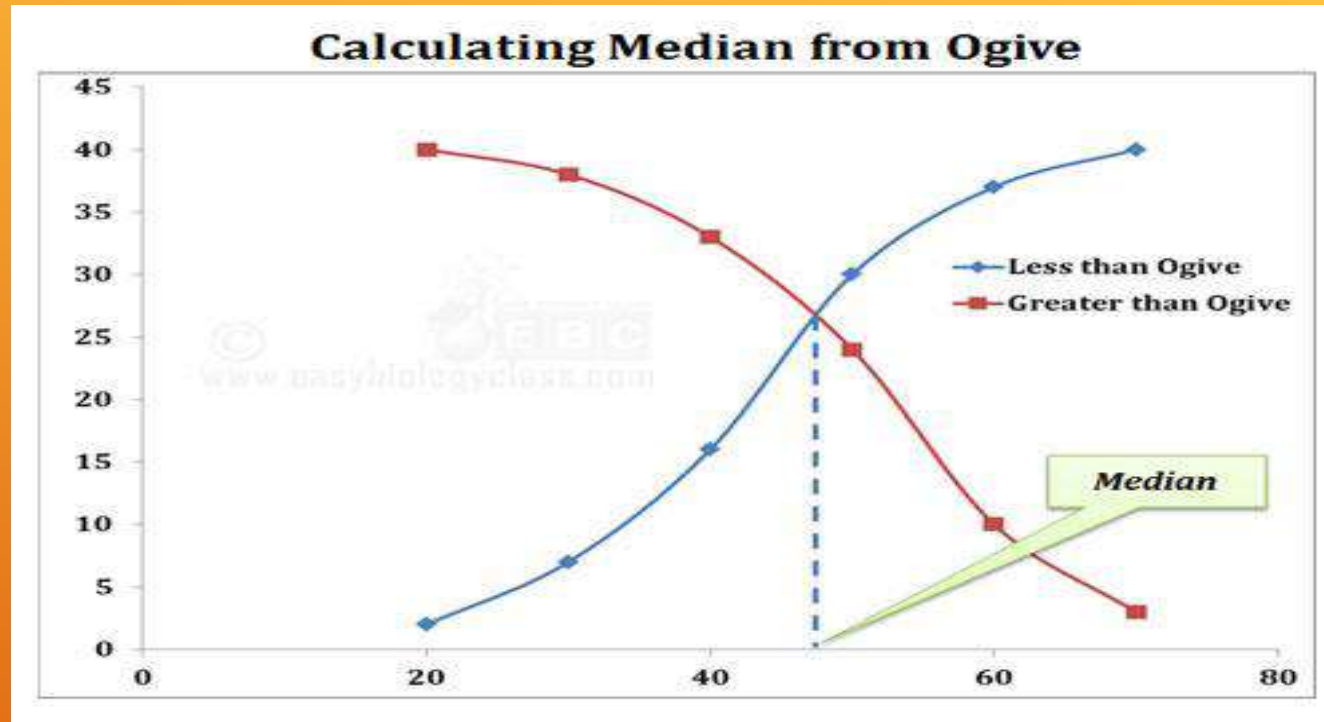
A frequency polygon is a graphical form of representation of data. It is used to depict the shape of the data and to depict trends. It is usually drawn with the help of a histogram but can be drawn without it as well.

Frequency Curve



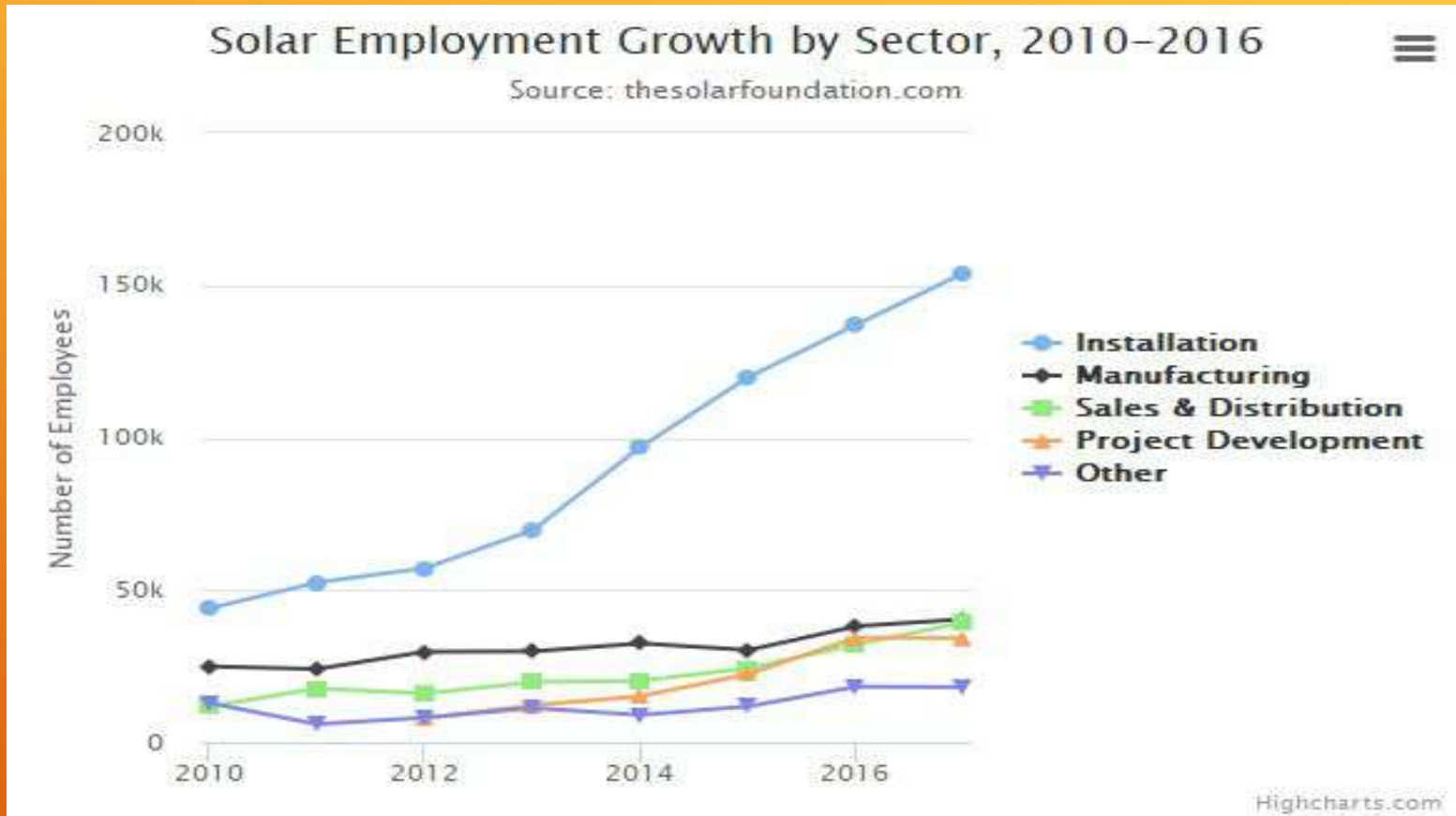
- It is a graphical representation of frequencies corresponding to their values by smooth curve.
- If the middle points of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that graph is called frequency curve
- It is limiting form of frequency polygon when the no. of observations become very large and class intervals are made smaller and smaller

Cumulative Frequency Curve



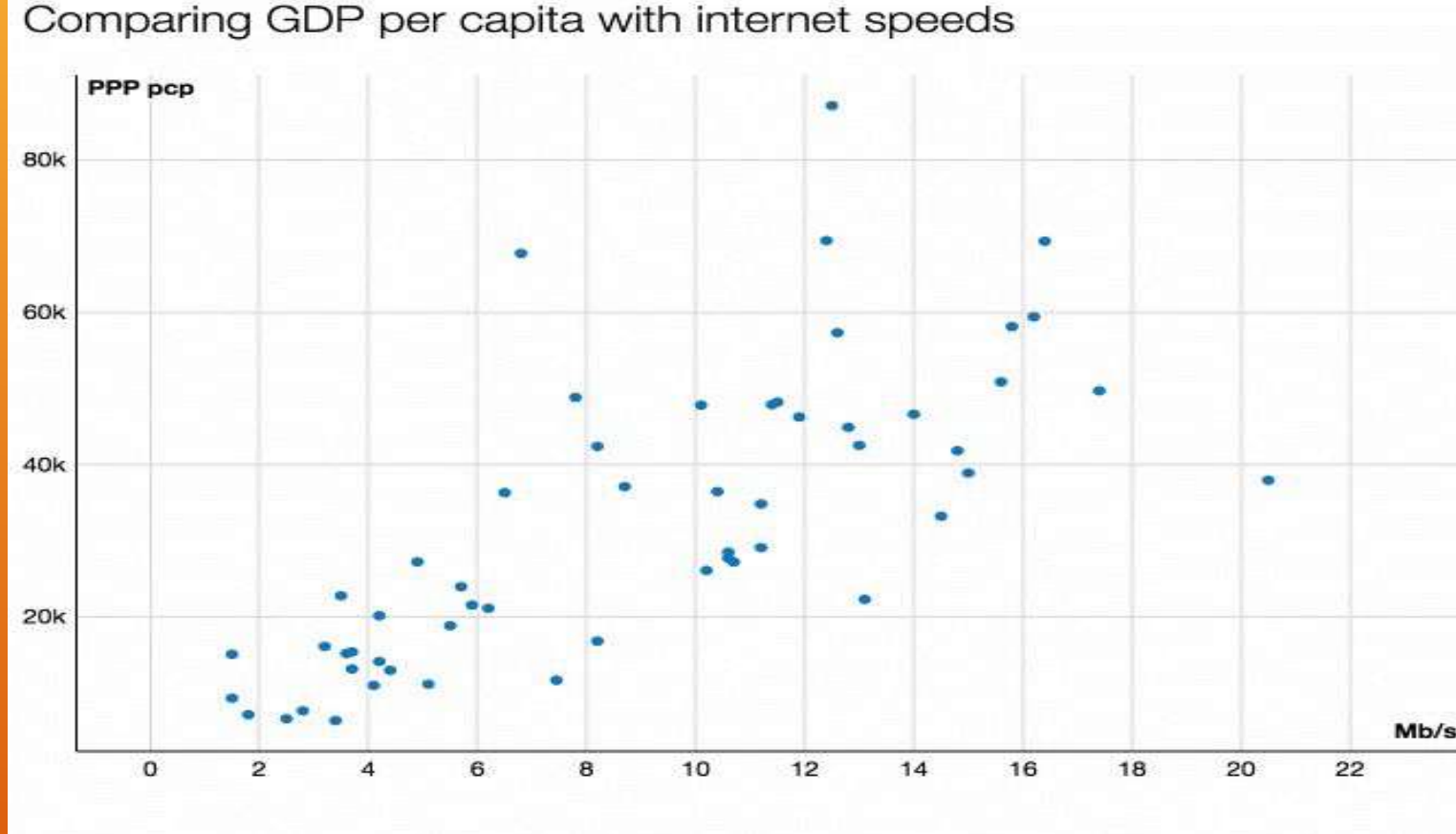
- It is graphical representation of the cumulative frequency distribution of continuous variable
- Points are plotted with c.f. along y-axis and the corresponding class boundaries along the x-axis and joining them freely
- It has two methods of constructing: less than ogive and more than ogive
- The intersect point of less than and more than ogive perpendicular to the x-axis gives the value of Median

Line Diagram



A **line chart** or **line graph** is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.

Scatter Plot



A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

Measure of Central Tendency

The central tendency of distribution is an estimate of the 'center' of a distribution of values or observation.

Objectives:

- Facilitate comparison between two or more groups
- To get a single representative value for whole series
- To know about population from sample
- To facilitate computation of other statistical measures
- To help in decision making
- To describe the distribution in concise manner

Requisites of a good measure of central tendency

- It should be rigidly defined (should have definite value)
- It should be based on all the observations
- It should not be affected by the extreme values
- It should be stable with regard of sampling
- It should be capable of further algebraic treatment
- It should be capable of being used in further statistical computations or processing

Various Measures of Central Tendency

1. Mathematical Averages

- Arithmetic mean: Simple, Weighted and Trimmed
- Geometric Mean
- Harmonic Mean

2. Positional Averages

- Median, Mode, Partition Values

Mean

Arithmetic mean:

- It is sum of the all observations divided by the no. of observations. It is the most common statistic of central tendency, and when someone says simply “the mean” or “the average,” this is what they mean.
- The arithmetic mean works well for values that fit the normal distribution i.e. homogeneous.
- Not work well for data that are highly skewed.

Sample Mean $\bar{x} = (\sum x_i) / n$ Population Mean $\mu = (\sum X_i) / N$

Geometric mean:

- If the variety values are measured as ratios, percentages, proportions GM gives a better measure of location than other means
- The geometric mean is the Nth root of the product of N observation.

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Harmonic mean:

- It is defined as the reciprocal of arithmetic average of the reciprocal of the given values
- It is suitable average when the data pertains to speed ratio, rates and velocity

The harmonic mean is the reciprocal of the arithmetic mean of reciprocals of the values.

$$\text{H.M.} = \frac{n}{\sum \frac{1}{x}}$$

Median:

- Median is the value which divide the distribution into two equal parts i.e. the number of observations in the first part is equal to the no. of observations in second part provided that the data is arranged either in ascending order or in descending order.
- The median is useful when you are dealing with highly skewed distributions.

Mean or Average

- **Arithmetic mean:**

$$\text{Mean} = \frac{\Sigma x}{n}$$

where Σ = Greek letter sigma, denotes 'sum of.'

n = number of values

- **For Discrete Series:**

$$\text{Mean} = \frac{\Sigma fx}{N}$$

where, f = frequency

- **For Continuous Series:**

$$\text{Mean} = A + \frac{\Sigma fd}{N} \times C$$

where $d = (X-A)/C$

A = Assumed Mean

C = Common divisor

Median

- **If the number of observations is odd:**

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

where n = number of observations

- **If the number of observations is even:**

$$\text{Median} = \frac{1}{2} \left[\left(\frac{n}{2}\right) \text{th term} + \left(\frac{n+1}{2}\right) \text{th term} \right]$$

- **For continuous series:**

$$\text{Median} = l + \frac{\left(\frac{N}{2} - c\right)}{f} \times h$$

where, l = lower limit of the median class

c = cumulative frequency of the preceding median class

f = frequency of the median class


h = class width

BASIS FOR COMPARISON	MEAN	MEDIAN
Meaning	Mean refers to the simple average of the given set of values or quantities.	Median is defined as the middle number in an ordered list of values.
What is it ?	It is an arithmetic average.	It is positional average.
Represents	Center of gravity of data set	Center of gravity of data set Mid-point of data set
Applicability	Normal distribution	Skewed distribution
Outliers	Mean is sensitive to outliers.	Median is not sensitive to outliers.
Calculation	Mean is calculated by adding up all the observations and then dividing the value obtained with the number of observations.	To calculate median, the data set is arranged in ascending or descending order, then the value that falls in the exact middle of the new data set, is median.

Mode

- The mode of a set of data is the most frequently occurring value.
- It requires that a continuous variable be grouped into a relatively small number of classes, either by making imprecise measurements or by grouping the data into classes.
- It is rarely useful to determine the mode of a set of observations, but it is useful to distinguish between unimodal, bimodal, etc. distributions, where it appears that the parametric frequency distribution underlying a set of observations has one peak, two peaks, etc.

Measure of Dispersion

- ▶ Averages are representatives of a frequency distribution but they fail to give a complete picture of the distribution.
 - ▶ They don't tell anything about the scatterness of observation within the distribution.
 - ▶ The scatterness or variation of observations from their average is known as dispersion.
 - ▶ There are different measures of dispersion like the range, the quartile deviation, the mean deviation and the standard deviation.
- 
- A series of four parallel diagonal lines in white and light orange, extending from the bottom right towards the center of the slide.

Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series.

Relative measures are not expressed in units but it is a pure number. It is the ratios of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.

► **Absolute Measures**

Range

Quartile Deviation

Mean Deviation

Standard Deviation

► **Relative Measure**

Co-efficient of Range

Co-efficient of Quartile Deviation

Co-efficient of mean Deviation

Co-efficient of Variation.

Range:

- This is simply the difference between the largest and smallest observations. This is the statistic of dispersion that people use in everyday conversation.
- Range is not very informative for statistical purposes. The range depends only on the largest and smallest values, so that two sets of data with very different distributions could have the same range, or two samples from the same population could have very different ranges, purely by chance.
- In addition, the range increases as the sample size increases; the more observations you make, the greater the chance that you'll sample a very large or very small value.

Range = Maximum – Minimum

Standard Deviation(S.D.):

-The standard deviation is the most widely used measure of variability.

-Standard Deviation is a measure that quantifies the degree of dispersion of the set of observations. The farther the data points from the mean value, the greater is the deviation within the data set, representing that data points are scattered over a wider range of values and vice versa.

-The standard deviation is the square root of the sum of the square deviation from mean divided by number of observations.

$$\text{Population S.D. } \sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

$$\text{Sample S.D. } s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}, \text{ where } (n-1) \text{ is known as degree of freedom (d.f.)}$$

- Variances the square of the standard deviation, and is important in various statistical calculations.

Coefficient of variation(C.V.)

- ▶ Coefficient of variation is the standard deviation divided by the mean; it summarizes the amount of variation as a percentage or proportion of the total.
- ▶ It is know as relative dispersion w.r.t. S.D. It is always expressed as %.
- ▶ It is useful when comparing the amount of variation for one variable among groups with different means, or among different measurement variables.
- ▶ The C.V. will be small if the variation is small. The one with less C.V. is said to be more consistent.
- ▶ **$C.V. = (S.D./Mean) \times 100$**

Thank You

