CA FOUNDATION - STATISTICS CHARTS BY - PROF. JATIN DEMBLA

STATISTICAL DESCRIPTION OF DATA

The word statistics has been derived either of the following Latin word "Status" which means a political State. Italian word "Statista" German word "Statistik" French word "Statistique

CHARACTERISTICS:
- Statistics are aggregate of facts.
- Statistics ate affected by a large number of causes
- Statistics are always numerically expressed.
- Statistics should be enumerated or estimated .
- Statistics should be collected in systematic manner.
- Statistics should be collected for pre-determined purpose.
- Statistics should be placed in relation to each other.

Important limitations
- Statistics does not deal with individual item.
- Statistics deals with quantitative data.
- Statistics laws are true only on averages.
- Statistics does not reveal the entire story.
- Statistics is liable to be misused.
- Statistics data should be uniform and homogeneous.

MODE OF PRESENTATION OF DATA
- Textual presentation
- Tabular presentation or Tabulation
- Diagrammatic presentation:

Types of diagrams
I. Line diagram
II. Bar diagram
III. Pie chart

Primary data	Secondary data
It's the data collected by a particular person or organization for his own use from the primary sources.	It is the data collected by some other person or organization for their own use but the investigator also gets it for his use.
1. Direct personal observation: In this method. The investigator collects the data personally and therefore, It gives reliable and correct information. USE: In case of natural calamity data can be collected more quickly and accurately by applying this method. 2. Indirect oral investigation In this method. A third person is contacted who is expected to know the necessary details about the persons for whom the enquiry is meant. USE: if there are some practical problems in reaching the respondents directly, as in case of rail accident. 3. Estimates from the local sources and correspondence: Here the investigator appoints agents and correspondents to collect the data. 4. Data through Questionnaire: The data can be collected by preparing a Questionnaire and getting it filled by the persons concerned. 5. Investigations through enumerators. This method generally employed by the Government for population census etc.	1. Information collected through newspapers and periodicals. 2. Information obtained from the publications of trade associations. 3. Information obtained from the research papers published by university departments or research bureaus or UG.C. 4. Information obtained from the official publications of the Central, State and the local governments dealing with crop statistics, Industrial statistics, Trade and transport statistics etc. 5. Information obtained from the official publications of the foreign governments for international organizations

MEASURES OF CENTRAL TENDENCY

Arithmetic mean or mean is the number which is obtained by adding the values of all the items of a series and dividing the total by the number of items.
When all items of a series are given equal importance than it is called simple arithmetic mean and when different items of a series are given different weights according with their relative importance is known weighted arithmetic mean.

Median is the middle value of the series when arranged in order of the magnitude. When a series is divided into more than two parts, the dividing values are called Partition values.

Quartiles are the measures which divide the data into four equal parts, each portion contains equal number of observation

There are three quartiles

If a statistical series is divided into four equal parts, the end value of each part is called a quartile and denoted by 'Q'.

The lower half of a data set is the set of all values that are to the left of the median value when the data has been put into increasing order.

The upper half of a data set is the set of all values that are to the right of the median value when the data has been put into increasing order.

The first quartile, denoted by Q1 is the median of the lower half of the data set. This means that about 25% of the numbers in the data set lie below Q1 and about 75% lie above Q1

The second quartile also called median and denoted by Q2 has 50% of the items below it and 50% of the items above it.

The third quartile, denoted by Q3, is the median of the upper half of the data set. This means that about 75% of the numbers in the data set lie below Q3 and about 25% lie above Q3

Formula of calculating mode in continuous series Mode =

$$L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

- Where L_1 = Lower limit of modal class
- F0 = Frequency of the group preceding the modal class
- F1 = Frequency of the modal class
- F2 = Frequency of the group succeeding the modal class
- C = Magnitude or class interval of the modal class

Deciles distribute the series into ten equal parts and generally expressed as D.
Percentiles divide the series into hundred equal parts and generally expressed as P.
Mode is the value which occurs most frequently in the series, that modal value has the highest frequency in the series.

Main purposes and functions of averages.
(i) To represent a brief picture of data.
(ii) Comparison.
(iii) Formulation of policies.
(iv) Basis of statistical analysis.
(v) One value for all the group or series.

Merits of Arithmetic mean:
- Simplicity
- Certainty
- Based on all values.
- Algebraic treatment possible.
- Basis of comparison
- Accuracy test possible
- No scope for estimated value

Merits of Median:
(i) Simple measure of central tendency.
(ii) It is not affected by extreme observations.
(iii) Possible even when data is incomplete.
(iv) Median can be determined by graphic presentation of data.
(v) It has a definite value.
(vi) Simple to calculate and understand
(vii) It is a positional value not a calculated value.

Merits of mode:
(i) Simple and popular measure of central tendency.
(ii) It can be located graphically with the help of histogram.
(iii) Less effect of marginal values.
(iv) No need of knowing all the items of series.
(v) It is the most representative value in the given series.
(vi) It is less effected by extreme values

Types of series	Direct Method	Shortcut Method	Step deviation method
Individual series	$\bar{X} = \frac{\sum x}{N}$	$\bar{X} = A + \frac{\sum d}{N}$	$\bar{X} = A + \frac{\sum d}{N} \times c$
Discrete Series	$\bar{X} = \frac{\sum fx}{N}$	$\bar{X} = A + \frac{\sum fd}{N}$	$\bar{X} = A + \frac{\sum fd}{N} \times c$
Continuous Series	$\bar{X} = \frac{\sum fm}{N}$	$\bar{X} = A + \frac{\sum fd}{N}$	$\bar{X} = A + \frac{\sum fd}{N} \times c$

Measure	Individual Series	Discrete Series	Continues Series	
	Size of item	Size of item	Size of item	Formula
Median	$\left(\frac{N+1}{2}\right)^{th}$ Item	$\left(\frac{N+1}{2}\right)^{th}$ Item	$\frac{N}{2}$	$l_1 + \left[\frac{\frac{N}{2} - cf}{f}\right] \times c$
First Quartile	$\left(\frac{N+1}{4}\right)^{th}$ Item	$\left(\frac{N+1}{4}\right)^{th}$ Item	$\frac{N}{4}$	$l_1 + \left[\frac{\frac{N}{4} - cf}{f}\right] \times c$
Third Quartile	$\frac{3(N+1)_{th}}{4}$ item	$\frac{3(N+1)_{th}}{4}$ item	$\frac{3(N)}{4}$	$l_1 + \left[\frac{\frac{3(N)}{4} - cf}{f}\right] \times c$

Combines Mean $\bar{x}_{12} = \frac{\bar{x}_1 n_1 \bar{x}_2 n_2}{N_1 + N_2}$

Weighted Mean $\bar{X} = \frac{\sum Wx}{\sum w}$

DISPERSION

Characteristics of Measures of Dispersion
- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations Based on all observations

Classification of Measures of Dispersion

The measure of dispersion is categorized as:
(i) An absolute measure of dispersion:
- The measures which express the scattering of observation in terms of distances i.e., range, quartile deviation.
- The measure which expresses the variations in terms of the average of deviations of observations like mean deviation and standard deviation.

(ii) A relative measure of dispersion:
We use a relative measure of dispersion for comparing distributions of two or more data set and for unit free comparison. They are the coefficient of range, the coefficient of mean deviation, the coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.

Variance of the Combined Series

If σ_1, σ_2 are two standard deviations of two series of sizes n_1 and n_2 with means $\bar{y}_1$ and $\bar{y}_2$. The variance of the two series of sizes $n_1 + n_2$ is:
$\sigma^2 = (1/n_1 + n_2) \div [n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)]$
where, $d_1 = \bar{y}_1 - \bar{y}$, $d_2 = \bar{y}_2 - \bar{y}$, and $\bar{y} = (n_1 \bar{y}_1 + n_2 \bar{y}_2) \div (n_1 + n_2)$.

Coefficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages. Also, when the unit of measurement is different. We need to calculate the coefficients of dispersion along with the measure of dispersion. The coefficients of dispersion (C.D.) based on different measures of dispersion are
- Based on Range = (X max – X min)/(X max + X min).
- C.D. based on quartile deviation = (Q3 – Q1)/(Q3 + Q1).
- Based on mean deviation = Mean deviation/average from which it is calculated.
- For Standard deviation = S.D./Mean

- Range - Range = $X_{max} - X_{min}$
- Quartile Deviation - Q = ½ × (Q3 – Q1)

Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \ldots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is
Mean deviation from average A = 1/n [$\sum_i |x_i - A|$]
For a grouped frequency, it is calculated as:
Mean deviation from average A = 1/N [$\sum_i f_i |x_i - A|$], N = $\sum f_i$
Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation. The standard deviation is given as
$\sigma = [(\sum_i (y_i - \bar{y})^2 / n]^{1/2} = [(\sum_i y_i^2 / n) - \bar{y}^2]^{1/2}$
For a grouped frequency distribution, it is
$\sigma = [(\sum_i f_i (y_i - \bar{y})^2 / N]^{1/2} = [(\sum_i f_i y_i^2 / n) - \bar{y}^2]^{1/2}$
The square of the standard deviation is the variance. It is also a measure of dispersion.
$\sigma^2 = [(\sum_i (y_i - \bar{y}) / n]^{1/2} = [(\sum_i y_i^2 / n) - \bar{y}^2]$
For a grouped frequency distribution, it is
$\sigma^2 = [(\sum_i f_i (y_i - \bar{y}) / N]^{1/2} = [(\sum_i f_i x_i^2 / n) - \bar{y}^2]$.
If instead of a mean, we choose any other arbitrary number, say A, the standard deviation becomes the root mean deviation.

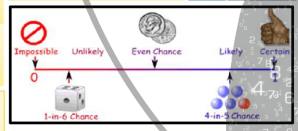
Coefficient of Variation

100 times the coefficient of dispersion based on standard deviation is the coefficient of variation (C.V.).
C.V. = 100 × (S.D. / Mean) = (σ/ȳ) × 100.

Correlation coefficient is the Statistic showing the degree of relation between two variables. The simple correlation coefficient, denoted r, ranges between -1 and +1 and quantifies the direction and strength of the linear association between the two variables. It is also called Pearson's correlation or product moment correlation coefficient. It measures the nature and strength between two variables of the quantitative type.

PROBABILITY

- The sum of all the probabilities in the sample space is 1.
- The probability of an event which cannot occur is 0.
- The probability of any event which is not in the sample space is zero.
- The probability of an event which must occur is 1.
- The probability of the sample space is 1.
- The probability of an event not occurring is one minus the probability of it occurring.
- The complement of an event E is denoted as E' and is written as P (E') = 1 - P (E)
- P (AUB) is written as P (A + B) and P (A ∩ B) is written as P (AB).
- If A and B are mutually exclusive events, P(A or B) = P (A) + P (B).
- When two events A and B are independent i.e. when event A has no effect on the probability of event B, the conditional probability of event B given event A is simply the probability of event B, that is P(B).
- If events A and B are not independent, then the probability of the intersection of A and B (the probability that both events occur) is defined by P (A and B) = P (A) P (B|A).



The probability distribution of a count variable X is said to be the binomial distribution with parameters n and abbreviated B (n,p) if it satisfies the following conditions:
- The total number of observations is fixed
- The observations are independent.
- Each outcome represents either a success or a failure.
- The probability of success i.e. p is same for every outcome.

- A and B are independent if P (B/A) = P(B) and P(A/B) = P(A).
- If E1, E2, ... En are n independent events then P (E1 ∩ E2 ∩ .. ∩ En) = P (E1) P (E2) P (E3)...P (En).
- Events E1, E2, E3, ... En will be pairwise independent if P(Ai ∩ Aj) = P(Ai) P(Aj) i ≠ j.
- P(Hi | A) = P(A | Hi) P(Hi) / ∑i P(A | Hi) P(Hi).
- If A1, A2, ...An are exhaustive events and S is the sample space, then A1 U A2 U A3 U ..... U An = S
- If E1, E2,....., En are mutually exclusive events, then P(E1 U E2 U ..... U En) = ∑P(Ei)
- If the events are not mutually exclusive then P (A or B) = P (A) +P (B) – P (A and B)
- Three events A, B and C are said to be mutually independent if P(A∩B) = P(A).P(B), P(B∩C) = P(B).P(C), P(A∩C) = P(A).P(C), P(A∩B∩C) = P(A).P(B).P(C)
- The concept of mutually exclusive events is set theoretic in nature while the concept of independent events is probabilistic in nature.
- If two events A and B are mutually exclusive
- P (A ∩ B) = 0 but P(A) P(B) ≠ 0 (In general)

⇒ P(A ∩ B) ≠ P(A) P(B)
⇒ Mutually exclusive events will not be independent.

Time series: A sequence of numerical data points in successive order, usually occurring in uniform intervals. A time series is simply a sequence of numbers collected at regular intervals over a period of time. Components of time series: The four components of time series are:
1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular variation

THEORETICAL DISTRIBUTION

Binomial Distribution:-

Mean = np
Variance = npq
- Variance is Always Less than Mean
- Note – Variance Will be Highest When P=Q=0.5
- Mode = (n+1)P (if Mode is Integer then it will be Bi-Modal & If Non Integer than Uni-Modal

Poisson Distribution:-

Mean = m
Varinace = m
Mode = As Same as Binomial Distribution

. Prob Of Success is Very Small
. It is Uniparametric Distribution

Normal Distribution Or Bell Curve Or Gaussian curve

. Mean, Median & Mode are Concide

. 1st & 3rd Quartile

a. Q1 = μ - 0.675 σ

b. Q3 = μ + 0.675 σ

2 Points of Inflexion

If a discrete random variable X has the following probability density function (p.d.f.), it is said to have a binomial distribution:
- $P(X = x) = nCx\ q^{(n-x)} p^x$, where q = 1 - p

p can be considered as the probability of a success, and q the probability of a failure.

Note: nCr ("n choose r") is more commonly written , but I shall use the former because it is easier to write on a computer. It means the number of ways of choosing r objects from a collection of n objects (see permutations and combinations).
If a random variable X has a binomial distribution, we write X ~ B(n, p) (~ means 'has distribution...').
n and p are known as the parameters of the distribution (n can be any integer greater than 0 and p can be any number between 0 and 1).
All random variables with a binomial distribution have may have different parameters (different values for n and p).

Expectation and Variance

If X ~ B(n,p), then the expectation and variance is given by:
- E(X) = np
- Var(X) = npq

Binomial & Normal Distribution Is Known As Biparametric Distribution

CORRELATION AND REGRESSION

Correlation

Correlation is a statistical method used to determine the extent to which two variables are related. Correlation analysis measures the degree of association between two or more variables. Parametric methods of correlation analysis assume that for any pair or set of values taken under a given set of conditions, variation in each of the variables is random and follows a normal distribution pattern. In scattered diagram, following elements are represented:
1. Rectangular coordinate
2. Two quantitative variables
3. One variable is called independent (X) and the second is called dependent (Y)
4. Points are not joined
5. No frequency table

In scatter plot, the pattern of data is indicative of the type of relationship between two variables. It may be
1. Positive relationship
2. Negative relationship
3. No relationship

The sign of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association.

Regression

Regression analysis encompass to identify the relationship between a dependent variable and one or more independent variables. Regression calculates the "best-fit" line for a certain set of data. The regression line makes the sum of the squares of the residuals smaller than for any other line. In regression analysis, a single dependent variable, Y, is considered to be a function of one or more independent variables, X1, X2, and so on. The values of both the dependent and independent variables are assumed as being determined in an error-free random manner.

In a simple regression analysis, one dependent variable is measured in relation to only one independent variable. The analysis is designed to develop an equation for the line that best models the relationship between the dependent and independent variables. This equation has the mathematical form:
Y = a + bx
In above equation, Y is the value of the dependent variable, X is the value of the independent variable is the intercept of the regression line on the Y axis when X = 0, and b is the slope of the regression line.

Correlation	Regression
• In correlation analysis the degree and direction of relationship between the variables are studies • If value of one variable is known, the value of other variable cannot be estimates • Correlation coefficient lies between -1 & 1 • Correlation coefficient is independent of change of origin and scale • With the help of correlation coefficient and standard deviation of two random variable (X,Y) regression coefficient can be obtained	• In regression analysis, the nature of relationship is studied • If value of variable is known, the value of other variable can be estimated using the functional relationships • Only one relation coefficient can be greater than 1 • Regression coefficient is independent opf change of origin but not of scale

INDEX NUMBER AND TIME SERIES

Index Numbers: An index number is an economic data figure reflecting price or quantity compared with a standard or base value. OF INDEX NUMBERS Index numbers are names after the activity they measure. Their types are as under :

Price Index : Measure changes in price over a specified period of time. It is basically the ratio of the price of a certain number of commodities at the present year as against base year.

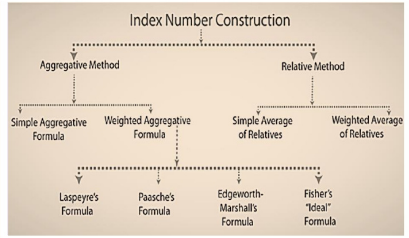
Quantity Index : As the name suggest, these indices pertain to measuring changes in volumes of commodities like goods produced or goods consumed, etc.

Value Index : These pertain to compare changes in the monetary value of imports, exports, production or consumption of commodities.

Simple index numbers: A simple index number Index numbers is a number that expresses the relative change in price, quantity, or value from one period to another. Let $p0$ be the base period price, and $p1$ be the price at the selected or given period. Thus, the simple price index is given by: $P = p1/p0\ (100)$

Weighted index
: *Weighted Index Numbers* = (∑*index number* × *weight*) /∑*weights*
There are two types of weighted indexes, they are
Lapeyre's index

1. According to the Fundamental Method $Q_{01} = \frac{q_1}{q_0} \times 100$
2. According to simple aggregative method $Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100$
3. According to simple relative method $Q_{01} = \sum \left[\frac{q_1}{q_0} \times 100\right]$ or $\frac{\sum 1}{n}$
4. According of the general weighted aggregative method $Q_{01} = \frac{\sum q_1 w}{\sum q_0 w} \times 100$
5. According to Laspeyre's aggregative method $Q_{01(p)} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$
6. According to Paasche's aggregative method $Q_{01(p)} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$
7. According to Drobish and Bowley's aggregative method $Q_{01} = \frac{1}{2}\left[\frac{\sum q_1 p_0}{\sum q_0 p_0} + \frac{\sum q_1 p_1}{\sum q_0 p_1}\right] \times 100$
8. According to Fisher's ideal method: $Q_{01(f)} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$
9. According to Fisher's ideal method: $Q_{01(M.E.)} = \frac{\sum q_1 p_0 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_0 p_1} \times 100$



Time – Series Component Analysis
- Used primarily for forecasting
- Observed value in time series in the aum or product of components
- Additive model Xt = Tt + St +Ct x Lt
- Multiplicative model (Linear in log Form) Xt = TtStCtLt