INTRODUCTION TO STATISTICAL METHODS OF ANALYSIS.

INTRODUCTION: -

Quantification is a mean to bring precision in any scientific enquiry. Its use in geography has given new dimensions to geographic research and has played a significant role in the development of the subject. Quantification is an approach in the observation of a phenomenon. We can observe a phenomenon qualitatively as well as quantitatively. Whereas research mostly concludes in qualitative terms, quantification helps us in arriving at these conclusions objectively.

There are two main approaches in Quantitative analysis; (1) Mathematical and (2) statistical.

(1) **Mathematical methods:** Mathematical methods includes those tools and techniques, like algebra, geometry, calculus and matrices etc., which are suitable to the phenomenon where the responses are exact in nature. Such responses are mainly found in the laboratory experiments of natural sciences. Since these responses are exact in nature a single experimental trial is sufficient to get the result. For example, if we are interested in finding out the specific heat of cupper (calories of heat require to raise the temperature one gram of the metal by one degree centigrade), we can carry out an experiment once and take the reading to get the result. Only one response is sufficient to get the result. There is no need of repeating this experiment due to the fact that any repetition will give the identical results, either today, tomorrow or at any other place.

(2) **Statistical Methods:** Statistical methods, which are based on principle aggregation, on the contrary are suitable to the phenomenon where the responses are not exact in nature. Such phenomenon is mainly found in social and behavioural sciences where human being are the respondents. In social sciences if we observe a phenomenon repeatedly, it may not give identical results. Every response will differ from the other marginally. Reason being that human responses do not work like machines of the laboratory. Human responses follow social laws which are superimposed by some behavioral elements also. Whereas the response of a social force may be identical for all the similar people in the society, the behavioral response varies from individual to individual, causing an element of error in the responses. No single response, therefore, give the accurate results and tools of mathematical methods of analysis do not help due to the presence of this element of error term. Statistical methods help us in finding out the underlying correct response by minimizing the error with the help of theory of aggregation.

THEORY OF AGGREGATION: -

As in human responses underlying true value is being superimposed by some element of error term, final responses will vary from individual to individual. Thus if the true value of an individual response is x_i super impose by some error term say ε_i, the final response X_i will be the sum of the two i.e.

$X_i = x_i + \varepsilon_I$

The error term ε_I is supposed to be caused by so many behavioral factors that cannot be predicted and hence is called as random error with the following assumptions:

- It is not large
- It operates both in positive as well as in negative directions
- Positive and negative values have equal probability.

With the above assumptions about the random error, it is possible to minimize it by the method of internal cancellation. If we take a sum of say 10 such responses, total of which will retain true response but some of the minus values of error tem will cancel the plus values internally. The sum total of these values will therefore carry only the balance amount of error. If we divide the sum by 10 the average value will be much closer to the true response than any individual response value. Aggregation will, thus, minimize the error effect substantially through internal cancellation.

The major role to be played in the theory of aggregation is the third assumption that both positive and negative values will operate with equal probability. This is possible only when we have larger amount of responses. For example, we know that probability of head and tail in a single throw of a coin is 0.5 i.e. if a coin is tossed n number of times, half of the time it will give head and half of the time tail. Does this mean that if coin is tossed twice we will get one head and one tail? Certainly not. It is quite possible that in both the cases we may get heads or tails. Then, what is the meaning of the probability? In fact, the probability starts working when the number of trails 'n' are large i.e. if we toss the coin large number of times then we will have half the times heads and half the times tails. The probability will become stronger as the size of the number of responses will increase. The same principal will apply to social responses also. If we have larger number of responses, the probability that half of them will be carrying positive random error and half will carry negative will become stronger and stronger as the size of response will increase.

When the quantitative responses carry the effect of random error, we can minimize the effect with the help of various statistical techniques using the principal of aggregation. We collect as many responses as are possible and extract the true value of the response by subjecting it to various

methods of statistical analysis depending on the problem. These **quantitative responses are known as data.** Thus for any statistical analysis following points will become evident:

(1) Data is the basic pre-requisite for any statistical analysis.
(2) For effective aggregation of data, the number of observations should be sufficiently large.
(3) Larger is the size of data, more reliable are the conclusions.
(4) Although there is no hard and fast rule about the minimum number of observations for any statistical analysis, generally, less than 7 or 8 number of observations is not desirable for any statistical analysis.
(5) For analyzing less than 7 or 8 number of observations, one should go for descriptive analysis rather than statistical.

DATA: -

Statistical methods relate to the techniques of extracting meaningful inferences from the real world situation manifested by different types of data. The data collected from the field is subjected to various types tools of analysis to drive the required inference. The required data is collected either through the primary sources or from the secondary sources. Secondary source of data is the data collected by different agencies for some general purposes and may serve only partially to any specific research. Data from the secondary sources may be Government reports, data published by official agencies such as telephone department, tax department, municipalities etc. A secondary data therefore seldom serves the purpose of any specific research. It is generally supplemented by the data collected by primary source from the field. For any field research the data is collected either by : (1) Census Enumeration Or through, (2) Sample Surveys. Researcher has to chose the method of data collection as per the resources at his/her disposal.

CENSUS ENUMERATION: -

When the data is collected for each and every observation included in the study area through an exhausted or complete enumeration it is known as "Census Enumeration". Census enumeration is a big exercise and requires considerable preparations. It requires a huge manpower to conduct a census and is equally costly also. Census enumerations are, therefore, conducted by the governments only in most of the cases. We have Population Census, Economic Census, Agricultural Census etc. in India with some fixed periodicity. Some time we also have even Census of Tigers and other endangered species for wild life conservation etc. Since in Census enumeration is complete enumeration of the study area, it is necessary that the area to be enumerated has to be demarcated completely which is known as the **Universe or the Population** of the study. The Universe or Population of any study will be decided by the research problem itself. For example, if the research study is concerned about house hold study of the tribal population living in Rajasthan,all the tribal households in Rajasthan will form the Universe of the study. However, if the study is limited to only to one tribe say "Bhils" of district of Jaisalmer, all the Bhil households living in Jaisalmer will form the universe. Similarly, Universe could be the agricultural farms of a district, or of a region. It could also be the workers in the textile factories of Maharashtra or sewage disposal plants of cities of a region etc. Questions are developed to be asked to the respondents written on a paper known as questionnaire. T team of enumerators goes to the field area of the study to get these questionnaires filled by asking the questions from the respondents. The answers are noted down on the questionnaire and become part of the data which is later subjected to various types of statistical analysis. All the statistics like mean, median, standard deviation, correlation, regression coefficients etc. related to Census data are known as **parameters** of the Universe.

SAMPLE SURVEYS: -

Census enumeration gives the true picture of any ground reality. If the ground for which data is to be collected is large, it will take longer time, will always be very expensive and will also require larger number of trained investigators and preparation for it will be quite difficult. Some time when researcher has only limited time and resources available, he or she will not be able to afford a heavy cost of Census enumeration and will also be not in a position to wait for long. In such situations most of the researchers prefer to go for a time saving technique of data collection, known as sample surveys, rather than Census enumeration. In a sample survey, instead of examining the ground situation for whole area only a group of some representative observations known as sample are selected with the hope that the real ground situation will be fairly well represented by the selected representative sample. The values like mean, median, standard deviation, correlation, regression coefficients etc. related to sample data are known as **statistics** of the samples. Powerful theory of sampling has been developed to draw inferences about the **parameters** of the Universe from these sample **statistics**. A properly selected sample saves considerable time and resources of the researchers. But, if the sample does not represent the universe, the whole exercise will go waste. It is therefore, important before the analysis of the sample data, to ensure that a sample drawn from a given universe is a proper representative of it. A sample can be drawn in several different ways. Which method of sampling will make it most representative of the Universe will depend on the structure of the Universe of the study.

NATURE OF STATISTICAL METHODS

- Statistical analysis starts with basic descriptive analysis of the given data, collected either through primary or through secondary source. To start with the data is available to the researcher is in a random order scattered on questionnaire. Before giving any serious dimension to the findings of the study it is important for the researcher to give introduction of the situation or area, which possible only when the data generated is transfer to a systematic and manageable form. Statistical methods of tabulation, graphical representation are helpful in this regards. Descriptive statistics also help us in visualizing the data in its integrated form with the help of its frequency curve and highlighting its differences quantitatively also with the descriptive statistics like central tendency, dispersion etc. At next level of statistical analysis, we learn as how to study the uncertainty with the help of the theory of "probability" and how to make comparison between two universes with the help of their sample statistics. We also learn to handle and measure the level of uncertainty in accepting or rejecting a hypothesis.

- Second stage of statistical analysis consist of the bivariate analysis in which me are exposed to the empirical methods of studying relationships between a dependent and an independent variable. The cause and effect relationship is further elaborated through bivariate regression analysis. Bivariate regression analysis is further extended to the analysis of multivariate regression analysis of one dependent variable and several independent variables.

- With the advent of high speed computers and soft were packages of various computer programmes of many important statistical technique, their application has become very simple. However, knowledge of matrix algebra is found to be very helpful in using these packages. A working knowledge of matrix algebra is therefore will also be an advantage to our students. Some modules on matrix algebra simplifying its uses in geography.

- **variables and constant**
- as data related to different **variables** is the basic input to any statistical analysis, it will be useful to understand the nature of different kinds of variables and the constant.
- **constants**
- constants as oppose to variables which do not change. there are certain characteristics whose value do not change even if we take several repeated observations such as length of a meter (100 cms.), minimum voting age of people of india (18 years), age of senior citizen (60 + years), atmospheric pressure at different points around an area etc. the characteristics whose values will not change at repeated observations (some of them as mentioned above) are known as constants.
- **variables**
- those characteristics whose values vary from observations to observations are known as variables. variables could be qualitative as well as quantitative.
- **qualitative variable**
- a qualitative variable is one which do not vary in terms of quantity, but varies in terms of its non- numeric attribute only. like: colours, red, yellow and green etc., gender: male and female, type of soil: black soil or alluvial soil and so. some of the qualitative variables could be in ordered form also, like; small and big, good and bad, developed and under developed and so on.
- **dummy variables**
- values of the qualitative are not amenable to numerical manipulation. some of the qualitative variables which are dichotomous in nature assuming the value of either zero or one are known as 'dummy variable' which can be subjected to limited mathematical manipulations also. a dummy variable will assume the value as ' 1' if the characteristic is present and '0' if it is absent. for a trichotomy also with some adjustment 'dummy' variables can be generated.
- **quantitative variables**
- quantitative variables are those whose different values are given in definite quantitative magnitude, like income of individuals, agricultural production, rainfall, number of floods in a year, number of hospitals in a region etc. quantitative variables are measured on interval or ratio scales both.
- quantitative variables could be **continuous as well as discontinuous**. continuous variables are those which can be measured in fractions also, like rainfall, area and productions of districts etc. a discontinuous variable is one whose values are found in integers only, like number of floods in a year. number of colleges, hospitals and towns etc. the variables could be **absolute variables as well as derived variables.**
- **absolute variable**

- An absolute variable is a variable whose values may relates to units of different sizes like several districts of different sizes. In such case even if there is no real correlation, the absolute values of these characteristics will show very high correlation as districts of larger areas will have larger values and districts of smaller areas will have smaller values of all these characteristics. This spurious correlation has to be controlled by describing the variable per unit of area through derived variables.

- **DERIVED VARIABLES OR INDICATORS.**

- Some other characteristics, however, which are standardized cannot be measured simply by an absolute variable for example; agricultural productivity, levels of literacy, levels of urbanization of an area, gender discrimination, force of mortality, level of fertility and human development etc. We can measure absolute agricultural production for districts of different sizes, but it will not indicate agricultural productivity of the area. As apart from the effect of soil fertility the districts of larger area will give larger agricultural production in comparison to the districts of smaller size. To remove the area effect, however, we can derive a new variable as an indicator of agricultural productivity of the soil through dividing agricultural production by size of the area i.e. agricultural productivity per unit of cropped area.

- Similarly, the other characteristics of complex nature can also be measured by deriving a new variable indicating the required characteristics. Such derived variables are known as indicators of different characteristics. Usually a statistical analysis is not carried out by taking the values of the absolute variables, instead we take the values of a carefully designed indicator.

- Derived variable are worked out by working out a ratio, rate or by a combination of different operations such as "Human Development Index" or 'Cost of Living Index Number' or by
  construction of any other complex composite index like "Principal Component Scores" etc.

- **INDICATORS DERIVED AS "RATIOS"**

- Whenever we take a variable for unit of observations of different sizes, we must remove the size effect by dividing the values by the size of the observations and work out the ratio of the value of the variable to the size or value of the variable per unit of size. Size could be the area, total population or a section of the population etc. In ratios both numerators and denominators are different. One does not include the other.

- Examples of some commonly used ratios.

- Density of population i.e. population per unit area.

- Agricultural productivity of land

- Agricultural productivity of labour

- Dependency ratio

- Urban rural ratios.

- Sex ratio.

- Child women ratio.

INDICATORS DERIVED AS "RATES" OR "PERCENTAGES"

Sometimes we observations are divided into sub categories and a researcher may be interested in the strength of each one of them in relation to total. In such a case we can derive the variables by taking the share of each one of them in proportion to the total. The proportion of each category can be worked out by dividing the value of each sub category by the total. Sum of the proportion of all sub categories will be automatically one. If we multiply each proportion, we get the percentage share of each sub category and their summation will be 100.

- In proportions or percentages, the value of the denominator includes the value of the numerator in such a way that sum of all the values of the numerator is equal to the total which goes to the denominator.
- Some of the common indicators of proportions or percentages are as given below:
- Percentage of urban population to total population.
- Percentage of area under different crops to total cropped area.
- Percentage of literate population to total population.
- Crude Birth Rate.
- Crude Death Rate.
- Population Growth Rate.
- Infant Mortality Rate.