**Chapter One**: **Introduction**

**1.1. Research Methods in Social Sciences**

Research is a scientific or critical investigation aimed at discovering facts and interpreting data or applying the evolved techniques in solving certain problems and give answers to research questions. It is an organized and systematic way of finding solutions to problems. It is systematic because there is a definite set of procedures and a series of steps which a researcher must follow. There are certain procedures in research process which must always be done in order to get the most accurate results.

Anybody who wishes to study any aspect of the world about them has to decide what methods they are going to use. Their decision is made on the basis of their assumptions about what kind of thing it is they are studying. Creswell (2003) clearly identified three approaches to social sciences research: Quantitative, Qualitative and Mixed methods approaches. The objective of this material is therefore to discuss in detail all about quantitative research techniques in social sciences such as Social Anthropology, sociology and development studies.

**1.2. The Natures of Quantitative Research**

*Quantitative**research*is an inquiry into an identified problem, based on testing a theory, measured with numbers, and analyzed using statistical techniques. The goal of quantitative methods is to determine whether the predictive generalizations of a theory hold true.

**1.3. Merits and Demerits of Quantitative Methods**

**Merits:**

 testing and validating already constructed theories about how and why phenomena occur;

 can generalize research findings when the data are based on random samples of sufficient size

 data collection using some quantitative methods is relatively quick (e.g., telephone interviews)

 data analysis is relatively less time consuming (using statistical software)

 the research results are relatively independent of the researcher (e.g., statistical significance)

 relatively more free of motivations, feelings, opinions, and attitudes of individuals who are carrying out the research

 it is useful for studying large numbers of people

 examines the relationships between and among variables critically.

 it seems more reliable method

**Demerits:**

Collect a much narrower and sometimes superficial dataset Results are limited as they provide numerical descriptions rather than detailed narrative and generally provide less elaborate accounts of human perception often carried out in an unnatural and artificial environment Preset answers will not necessarily reflect how people really feel about a subject. The researcher’s categories that are used might not reflect local constituencies’ understandings Knowledge produced might be too abstract and general for direct application to specific local situations, contexts, and individuals.

**1.4. Underlying Assumptions of Quantitative Methods**

 Reality is objective, “out there,” and independent of the researcher – therefore reality is something that can be studied objectively;

 The values of the researcher do not interfere with, or become part of, the research

 is value-free; the researcher should remain distant and independent of what is being researched;

 Research is based primarily on deductive forms of logic and theories and hypotheses are tested in a cause-effect order; and

**1.5 Theoretical Backup of Quantitative Research**

**Positivism**

Positivism is a philosophical concept, and refers to a particular set of assumptions about the world and about appropriate ways of studying it. In general, positivists see ‘society’ as more important than the ‘individual’. For example, they point out that individuals are born, take their place in society and then die, but society continues largely undisturbed. Moreover, positivists suggest that people are the puppets of society, i.e. they are controlled by social forces emanating from the organization of society. This is because they believe that just as there are natural laws governing the behavior of chemicals, elements, plants, animals, etc., so there are social forces or laws governing and determining the operations of the social world, particularly our everyday experiences and life chances. Such laws are the product of the way a society or social group is socially organized, i.e. its social structure, and are beyond human influence. Both functionalism and Marxism are positivist theories because they believe that individual behavior is less important for our understanding of social life than the social structure of society. Functionalism stresses the need for individuals to be socialized into a value consensus which shapes and controls the behavior of members of society, and brings about social order. It is consequently very difficult to resist social pressures to conform to certain values and norms such as achieving in education, working for a living, etc. Marxism is also positivist because it sees human behavior as shaped or determined by the economic organization of capitalist society. Marxists argue that capitalist societies like the UK are characterized by profound class inequalities in wealth and income distribution, education, health, mortality, etc. In particular, Marxists argue that our behavior is a product of our socio-economic positions within capitalist society.

In other words, the social class to which we belong exerts a strong influence on our life chances and outcomes, e.g. whether or not we live to a ripe old age, what causes our death, what standard of living we experience and so on. Positivists see sociology as the ‘science of society’ and believe that the behavior of human beings can be objectively and scientifically measured in much the same way as the subject matter of the natural sciences. They consequently argue that sociologists should adopt the logic and methods of the natural sciences in their exploration of how the social structure of society shapes people’s behavior and actions. This approach, they argue, will produce scientific laws of human behavior.

If we examine positivist principles further, we can see certain assumptions about the characteristics that scientific method should ideally have. First, research should be objective or value-free. In other words, the sociologist should be neutral and not allow their personal or political opinions and prejudices to bias any aspect of their research method or their interpretation of the data they collect. The sociologist should be determined to pursue scientific truths with an open mind. One way in which to ensure objectivity is to carry out research under controlled conditions. Natural scientists have the advantage of conducting experiments in laboratories. However, sociologists very rarely use laboratory experiments and have had to devise alternative methods of control. As we shall see, positivists have developed sampling techniques and rules of questionnaire design, as well as keeping at a distance from the people whom they are studying, in order to maintain what they regard as objective control.

Second, positivists regard reliability as the most important characteristic of scientific method. They argue that the sociological research method used in any piece of research should be able to be repeated by other sociologists in order to verify and check its scientific accuracy. The research method should be open to inspection, criticism and testing by other researchers. Positivists, as we shall see, regard research methods that produce quantitative data as more reliable than other methods because they are normally organized in standardized and systematic ways, e.g. a logical sequence of set questions involving tick-boxes is easily replicated by another sociologist.

Third, the research should produce mainly quantitative or statistical data that can be converted into tabular or graphical information. As we shall see, some primary research methods, notably the survey questionnaire and the structured interview are most likely to produce this type of data.

As we have seen, positivists tend to believe that the causes of human behavior lie *outside* of the individual in the structural forces of society. Consequently, they tend to take a ‘macro’ approach to the study of society in that they are primarily concerned with examining the relationships between different parts of the social structure, e.g. the impact of the economy on education, rather than how individuals see the social world. They therefore see little point in employing qualitative research methods that attempt to see the world through the eyes of individuals such as participant observation and unstructured interviews.



**1.6. Types of Quantitative Research**

Three general types of quantitative methods:

1. **Experimental Research Design**

Experimental research answers the question “What if?” The aim of the experimental research is to investigate the possible cause-and-effect relationship by manipulating one independent variable to influence the other variable(s) in the experimental group, and by controlling the other relevant variables, and measuring the effects of the manipulation by some statistical means. By manipulating the independent variable, the researcher can see if the treatment makes a difference on the subjects. True experiments are characterized by random assignment of subjects to experimental conditions and the use of experimental controls.

If the average scores of two groups prove to be significantly different, and if there are not any explanations for this difference, then it can be concluded that the effect of the treatment caused this difference. This is where experimental research differs from correlation research, For instance, correlation studies only describe or predict the strong *relationship* between socioeconomic level and the academic achievement but cannot prove the direct cause-and-effect relationship between these two variables. It is the experimental research which can demonstrate that by changing the independent variable, a change is possible on the dependent variable.

In educational research the most frequently studied dependent variables are achievement, motivation, attention, interest in learning, participation and attitudes. The common independent variables that are manipulated are teaching methods, types of assignments, types of teaching materials such as text books and visual aids, types of rewards, types of questions used by the teacher, and evaluation techniques. There are however, some independent variables such as age and gender that cannot be manipulated. When the independent variable that is chosen cannot be manipulated, either a comparative research is conducted, or a second independent variable is chosen for manipulation in order to conduct an experimental study.

To investigate the effects of a new drug abuse prevention program on the attitudes of junior high school students using experimental and control groups who are either exposed or not exposed to the program respectively, and using a pretest-posttest design in which only half of the students randomly receive the pretest to determine how much of an attitude change can be attributed to pretesting or the educational program.

1. **Quasi-Experimental Design**

Quasi-experimental studies share almost all the features of experimental designs *except* that they involve non-randomized assignment of subjects to experimental conditions.

1. **Surveys Design**

Surveys include **cross-sectional** and **longitudinal** studies using questionnaires or interviews for data collection with the intent of estimating the characteristics of a large population of interest based on a smaller sample from that population.

**Chapter Two: Survey Research**

**2.1. Survey Research Design**

We live in an information society. There is an ever-growing demand for statistical information about the economic, social, political, and cultural shape of countries. Such information will enable policy makers and others to make informed decisions for a better future. Sometimes, such statistical information can be retrieved from existing sources, for example, administrative records. More often, there is a lack of such sources. Then, a survey is a powerful instrument to collect new statistical information.

A social survey is a method of obtaining large amounts of data, usually in a statistical form, from a large number of people in a relatively short time. It usually takes the form of a self-completion questionnaire (this may be handed to the respondent or sent through the post) or an interviewer may read the questions to the respondent and fill in the questionnaire (otherwise known as an ‘interview schedule’) on behalf of the respondent.

The survey method is regarded as scientific because surveys are normally carried out under controlled conditions. They are organized in a logical and systematic fashion via questionnaire design. If the research is properly carried out, the personal influence of the researcher on the results is slight. In other words, they are seen to be objective and value-free. Very importantly, they are regarded as highly reliable because they are easily replicated and the quantifiable data can be verified by others. Moreover, the survey produces large amounts of statistical information, relatively quickly and cheaply, which enables comparisons to be made between different groups and populations.

A survey collects information about a well-defined population. This population need not necessarily consist of persons. For example, the elements of the population can be households, farms, companies, or schools. Typically, information is collected by asking questions to the representatives of the elements in the population. To do this in a uniform and consistent way, a questionnaire is used. One way to obtain information about a population is to collect data about all its elements. Such an investigation is called a census or complete enumeration. This approach has a number of disadvantages:

1. It is very expensive. Investigating a large population involves a lot of people (e.g., interviewers) and other resources.

2. It is very time-consuming. Collecting and processing a large amount of data takes time.

3. Large investigations increase the response burden on people.

A survey is a solution to many of the problems of a census. Surveys collect information on only a small part of the population. This small part is called the sample.

In principle, the sample provides information only on the sampled elements of the population. No information will be obtained on the non sampled elements. Still, if the sample is selected in a “clever” way, it is possible to make inference about the population as a whole. In this context, “clever” means that the sample is selected using probability sampling. A random selection procedure uses an element of chance to determine which elements are selected, and which are not.

**2.2. Characteristics of the Survey Method**

The following are the main characteristics of the survey method of research:

 The survey method gathers data from a relatively large number of cases at a particular time.

 It is essentially cross-sectional.

 It is not concerned with the characteristics of individuals.

 It involves clearly defined problem.

 It requires experts imaginative planning.

 It involves definite objectives.

 It requires careful analysis and interpretation of the data gathered.

 It provides information ‘useful to the solution of local problems.

**2.3. The Advantages and Limitations of Survey Research**

**1. Advantages of Surveys**

Surveys are primarily used to collect quantitative information on the perceptions and opinions of a sample of people which adequately represent the population of interest. They are appropriate for measuring people’s perceptions, opinions, knowledge, attitudes, behavioral intentions, and behavior using primarily closed-ended questions.

1. Can complete structured questions with many stakeholders within a relatively short time frame.

2. Can be completed by telephone, mail, fax, or in-person.

3. It is quantifiable and generalisable to an entire population if the population is sampled appropriately.

4. Standardized, structured questionnaire minimizes interviewer bias.

5. Tremendous volume of information can be collected in short period of time.

6. Can take less time to analyze than qualitative data.

Surveys *should not* be used to:

1. Replace qualitative techniques

2. Explore opinions in-depth

**2. Disadvantages of Surveys**

 The major disadvantages of the questionnaire are the possibility of the misinterpretation of the questions. Misinterpretations are due to the respondent’s willingness or impersonality.

 More difficult to collect a comprehensive understanding of respondents’ perspective (in-depth information) compared to in-depth interviews or focus groups.

 Requires some statistical knowledge, sampling and other specialized skills to process and interpret results.

o The first problem is faced in planning a questionnaire and its development.

o Another problem is to get adequate answer or information through questionnaire. Return of questionnaire is always doubtful.

o validity of the data through questionnaire is doubtful.

o Selection of representative sample.

**2.4. Stages in a survey**

**Stages in a survey research are:**

1. Choosing the topic to be studied

2. Reviewing the literature

3. Forming of hunches and hypotheses

4. Identifying the population to be surveyed

5. Carrying out preparatory investigations and interviews

6. Drafting the questionnaire or interview schedule

7. Conducting a pilot survey

8. Finalizing the questionnaire

9. Selecting a sample of the population

10. Selecting and training interviewers (if necessary)

11. Collecting the data

12. Processing the data and analyzing the results

13. Writing the research report, perhaps in the form of a book

14. Publication of the report.

**2.5. Relevant Concepts in Survey Research**

**2.5.1. Hypotheses**

Hypo + thesis = Hypothesis

‘Hypo’ means tentative or subject to the verification and ‘Thesis’ means statement about solution of a problem. The hypothesis is a tentative solution of a problem.

Hypothesis offers a solution of the problem that is to be verified empirically and based on some rationale. It is a statement temporarily accepted as true in the light of what is known at the time about the phenomena. It is the basis for planning and action- in the research for new truth. It is also known as tentative theory, after verification it takes the shape of final theory. A theory embers new hypotheses, these are subjected to verification, after the verification it becomes a new theory in field studies. In building up the theories, this cyclic process continues.

**Importance of a Hypothesis**

 *Hypothesis as the Investigator*’*s* “*Eyes*”: Carter V. Good thinks that by guiding the investigator in further investigation it serves as the investigator’s “Eyes” in seeking answers to tentatively adopted generalization.

 *It Focuses Research*: Without it, research is unfocussed research and remains like a random empirical wandering. It serves as necessary link between theory and the investigation.

 *It Places Clear and Specific Goals*: A well thought out set of hypothesis is that they place clear and specific goals before the research worker and provide him with a basis for selecting sample and research procedure to meet these goals.

 *It Prevents Blind Research*: “The use of hypothesis prevents a blind search and indiscriminate gathering of masses of data which may later prove irrelevant to the problem under study.”

George J. Motley thinks that Hypotheses serve the following purposes:

 They provide direction to research and prevent the review of irrelevant literature and the collection of useful or excess data. A hypothesis is a tentative assumption drawn from knowledge and theory which is used as a guide in the investigation of other facts and theory that are as yet unknown.

**Sources of Hypotheses**

These sources are namely theoretical background, knowledge, insight and imagination that come from instructional program and wide reading experiences, familiarity with existing practices.

* + 1. **Variables in a Hypothesis**

A variable is a factor or character that can have more than one or different values such as yield, height, weight and family size. The idea behind a variable is that it can have different values, that characteristics of objects, events or people can be measured along some continuum that forms a uniform numerical scale. For example, age (in years) and

These variables normally start out as concepts, coming from either research questions or hypotheses. First, it is necessary to define the concept in terms of the meaning it is to have in a particular research project. For example, age might be defined as ‘years since birth’, and education as ‘the highest level of formal qualification obtained’.

The second step is to *operationalize* the concept to show how data related to it will be generated. This requires the specification of the procedures that will be used to classify or measure the phenomenon being investigated. The way a concept is defined and measured has important consequences for the kinds of data analysis that can be undertaken. A hypothesis is made testable by providing operational definitions for the terms or variables of the hypothesis.

**The Independent Variable:** The independent variable which is a stimulus variable or input operates either within a person or within environment to affect his behavior. It is that factor which is measured, manipulated or selected by the experimenter to determine its relationship to an observed phenomenon.

If a researcher is studying the relationship between two variables X and Y, and if X is independent variable, then it affects another variable Y. It is the cause for change in other variables.

**The Moderator Variable:** The term moderator variable describes a special type of independent variable a secondary independent variable selected for study to determine if it affects the relationship between the primary independent variable and the dependent variable. The moderator variable is defined as that factor which is measured, manipulated or selected by the experimenter to discover whether it modifies the relationship of independent variable to an observed phenomenon.

**Control Variable:** All the variables in a situation cannot be studied at the same time; some must be neutralized to guarantee that they will not have a differential or moderating effect on the relationship between the independent and dependent variables. These variables whose effects must be neutralized or controlled are known as control variables. They are defined as those factors which are controlled by experimenter to cancel out or neutralize any effect they might otherwise have on the observed phenomena.

Teachers given more positive feedback-experiences will have more positive attitudes towards children than teachers given fewer positive feedback-experiences. From this hypothesis:

**Independent Variable**: Number of positive feedback experiences for teacher.

**Dependent Variable**: Possessiveness of teacher’s attitude towards students.

We may hypothesize that climate and soil will have beneficial effects on crop productivity. In this situation climate and soil are the Independent Variables (IV) and crop productivity is the Dependent Variable (DV). However, it is very likely that the DV (the amount of crop production per unit area) may be suffering from other factors such as type of seed and fertilizers. All these other potential sources of influence are known as extraneous variables.

* + 1. **Levels of Measurement**

In quantitative research, aspects of social reality are transformed into numbers in different ways. *Measurement* is achieved either by the assignment of objects, events or people to discrete categories, or by the identification of their characteristics on a numerical scale, according to arbitrary rules. The former is referred to here as *categorical* measurement and the latter as *metric n* measurement.

**Nominal-level measurement**

In *nominal-level measurement*, the categories must be homogeneous, mutually exclusive and exhaustive. This means that all objects, events or people allocated to a particular category must share the same characteristics, they can only be allocated to one category, and all of them can be allocated to some category in the set. The categories have no intrinsic order to them, as is the case for the categories of gender or religion. People can also be assigned numbers arbitrarily according to some criterion, such as different categories of eye color – blue (1), brown (2), green (3), etc. However, these categories have no intrinsic order (except, of course, on the color spectrum).

**Ordinal-level measurement**

The same conditions apply in *ordinal-level measurement*, with the addition that the categories *are* ordered along some continuum. For example, people can be assigned numbers in terms of the order in which they cross the finishing line in a race, they can be assigned social class categories (‘upper’, ‘middle’ and ‘lower’) according to their income or occupational status, or they can be assigned to age categories (‘old’, ‘middle-aged’ and ‘young’) according to some criterion.

However, the intervals between such ordinal categories need not be equal. For example, the response categories of ‘often’ (1), ‘occasionally’ (2) and ‘never’ (3) cannot be assumed to be equally spaced by researchers, because it cannot be assumed that respondents regard them this way. When the numbers in brackets are assigned to these categories, they only indicate the order in the sequence, not how much of a difference there is between these categories.

Similarly, the commonly used Likert categories for responses to attitude statements, ‘strongly agree’, ‘agree’, ‘neither agree nor disagree’, disagree’, and ‘strongly disagree’, are not necessarily evenly spaced along this level of agreement continuum, although researchers frequently assume that they are.

**Metric Measurement**

The introduction of the simple idea of equal or measurable intervals between positions on a continuum transforms categorical measurement into *metric measurement*.

***Interval-level measurement*** is achieved when the categories or scores on a scale are the same distance apart. Whereas in ordinal-level measurement the numbers ‘1’, ‘2’ and ‘3’ only indicate relative position, say in finishing a race, in interval-level measurement, the numbers are assumed to be the same distance apart – the interval between ‘1’ and ‘2’ is the same as the interval between ‘2’ and ‘3’.

The distinguishing feature of interval-level measurement is that the zero is arbitrary. Whatever is being measured cannot have a meaningful zero value. In this case, a zero score is achieved by an arbitrary decision about what numbers to assign to the response categories. It makes no sense to speak of a zero attitude, only relatively more positive or negative attitudes.

**Ratio-level measurement**

*Ratio-level measurement* is the same as interval-level measurement except that it has an absolute or true zero. For example, goals scored in football, or age in years, both have absolute or true zeros; it is possible for a team to score no goals, and a person’s age is normally calculated from the time of birth – point zero.

Ratio-level measurement is not common in the social sciences and is limited to examples such as age (in years), education (in years) and income (in dollars or other currencies). This level of measurement has only a few advantages over the interval level of measurement, mainly that statements such as ‘double’ or ‘half ’ can be made. For example, we can say that a person aged 60 years is twice as old as a person aged 30 years, or that an income of $20,000 is only half that of $40,000. These kinds of statements cannot be made with interval-level variables. For example, with attitude scales, such as those discussed above, it is not legitimate to say that one score (say 40) is twice as positive as another (say 20).

In most social science research, this limitation is not critical; interval-level measurement is usually adequate for most sophisticated forms of analysis. However, we need to be aware of the limitations and avoid drawing illegitimate conclusions from interval-level data.

Table 1.2

|  |  |  |  |
| --- | --- | --- | --- |
| *Levels of*  *measurement* |  |  |  |
| **Level** | **Description** | **Types of categories Examples** | **Examples** |
| Nominal | A set of categories for classifying objects, events or people, with no assumptions about order | Categories are homogeneous, mutually exclusive and exhaustive. | Marital status  Religion Ethnicity |
| Ordinal | As for nominal level measurement, except the categories are ordered from highest to lowest. | Categories lie along a continuum but the distances between them cannot be assumed to be equal. | Frequency (often,  sometimes, never)  Likert scale |
| Interval | A set of ordered and equal- interval categories on a contrived measurement scale. | Categories may be discrete or continuous with arbitrary intervals and zero point. | Attitude score IQ  score Celsius scale |
| Ratio | As for interval-level  measurement | Categories may be discrete or continuous but with an absolute zero. | Age Income No. of  children |

**Transformations between Levels of Measurement**

It is possible to transform metric data into categorical data but, in general, not the reverse. For example, in an attitude scale, scores can be divided into a number of ranges (e.g. 10–19, 20–29, 30–39, 40–50) and labels applied to these categories (e.g. ‘low’, ‘moderate’, ‘high’ and ‘very high’). While such transformations may be useful for understanding particular variables, and relationships between variables, measurement precision is lost in the process, and the types of analysis that can be applied are reduced in sophistication.

The reason why careful attention must be given to level of measurement in quantitative research is that the choice of level determines the methods of analysis that can be undertaken. In designing a research project, decisions about the level of measurement should be used for each variable need to anticipate the type of analysis that will be required to answer the relevant research question(s).

**Chapter Three: Sampling**

***3.1. Population/Universe/Sampling Frame***

In order to apply a sampling technique, it is necessary to define the *population* (also called the target population, universe or sampling frame) from which the sample is to be drawn. A population is any entire collection of people, animals, plants or things from which we may select sample data. It is the entire group we are interested in, which we wish to describe or draw conclusions about. In order to make any generalizations about a population, a sample, that is meant to be representative of the population, is often studied. A sample statistic gives information about the corresponding population parameter. Sampling frame is the actual set of units from which a sample has been drawn. In other words, it is the listing or displaying of all accessible population from which the researcher draws his/her samples. It is a common practice in social research to work with samples rather than populations, particularly when the population under consideration is very large, such as a country, a region or an urban area. It is tedious and expensive to study such large populations. However, there is a price to pay when samples are used: first, it is not possible to be absolutely confident that what a sample shows also exists in the population; second, it adds considerable complication to the analysis.

***Samples***

A sample is a group of items selected from a larger group (the population/universe) for any statistical analysis. It is assumed that the sample is the utmost perfect representative of the general population. A sample is generally selected for study because the population may be too large to study in its entirety; or it may be too costly and time consuming to deal with each and every population under consideration.

Before selecting the samples, it is important that the researcher must carefully and completely defines the population, including a description of the members to be included.

Sampling is used for a variety of reasons. Studying a whole population may be slow and tedious; it can be expensive and is sometimes impossible; it may also be unnecessary. Given limited resources, sampling can not only reduce the costs of a study, but also, given a fixed budget, increase the breadth of coverage. However, as mentioned earlier, there is a price to pay when it comes to analyzing the data.

**Characteristics of a Good Sample**

The following are the main characteristics of a good sample:

 A good sample is the **true representative** of the population corresponding to its properties.

 A good sample is **free from bias**, the sample does not permit prejudices the learning and preconception, imaginations of the investigator to influence its choice.

 A good sample is an objective one, it refers objectivity in selecting procedure or absence of subjective elements from the situation.

 A good sample maintains accuracy. It yields an accurate estimates or statistics and does not involve errors.

 A good sample is **comprehensive in nature**. This feature of a sample is closely linked with true-representativeness. Comprehensiveness is a quality of a sample which is controlled by specific purpose of the investigation. A sample may be comprehensive in traits but may not be a good representative of the population.

A good sample is also **economical** from energy, time and money point of view.

 The subjects of good sample are easily approachable. The research tools can be administered on them and data can be collected easily.

 The size of good sample is such that it yields an accurate result. The probability of error can be estimated.

**Stages in Sampling Process**

Most literatures recommend the under-mentioned stages while sampling to collect primary data.

 Defining the concerned population

 Specifying a sampling frame, a set of items or events possible to measure

 Specifying a sampling method for selecting items or events from the frame

 Determining the sample size

**Sampling Methods**

**Probability Samples**

In a *probability sample*, every population element must have a known and non-zero chance of being selected. Most types of probability sample will also give every element an equal chance of being selected. A *non-probability sample* does *not* give every population element a chance of selection.

To be able to use the results obtained from a sample to draw conclusions about a population, the sample must be selected using probability techniques.

**Characteristics of Probability Sampling**

The following are the main characteristics of probability sampling:

In probability sampling we refer from the sample as well as the population.

 In probability sampling every individual of the population has equal probability to be taken into the sample.

 Probability sample may be representative of the population.

 The observations (data) of the probability sample are used for the inferential purpose.

 The probability is comprehensive

**Characteristics of Non-probability Sampling**

 There is no idea of population in non-probability sampling.

 There is no probability of selecting any individual.

 The observations of non-probability sample are not used for generalization purpose.

 Non-parametric or non-inferential statistics are used in non probability sample.

**Types of Probability Sampling**

***1. Simple random sampling***

The first, *simple random sampling*, involves a selection process that gives every possible sample of a particular size the same chance of selection. Each element of a population must be able to be identified and numbered. Once the sample size has been decided, numbers are selected using a table of random digits until the desired total is reached.

A simple random sample is obtained when all the units in the reference population have the *same probability* of being included in the sample. In order to implement this sampling design, the researcher will, first of all, need a complete list of the members of the population; a number will then be assigned to each of the *N* units in the list, and the *n* numbers corresponding to the subjects for inclusion in the sample will be picked out at random. is done by using a number of techniques as :

(*a*) Tossing a coin.

(*b*) Throwing a dice.

(*c*) Lottery method.

(*d*) Blind folded method.

(*e*) By using random table of ‘Tippett’s Table’.

**Advantages**

o It requires a minimum knowledge of population.

o It is free from subjectivity and free from personal error.

o It provides appropriate data for our purpose.

o The observations of the sample can be used for inferential purpose.

**Disadvantages**

o The representativeness of a sample cannot be ensured by this method.

o This method does not use the knowledge about the population.

o The inferential accuracy of the finding depends upon the size of the sample.

**2. *Systematic sampling***

*Systematic sampling* – A procedure that is statistically equivalent to simple random sampling – from the point of view of the result (in the sense that it produces a simple random sample) – is that of *systematic sampling*. The only difference lies in the technique of picking out the subjects. The sampling units are no longer selected by lottery (or random number tables), but by scrolling the list of subjects and systematically selecting one unit at every given interval. If the size *N* of the reference population is known and the size *n* of the sample has been established, one unit every *k* =*N/n* units of the population is selected, beginning with a number chosen at random between 1 and *k* (*k* is called the *sampling interval*). In social research, systematic sampling is often used precisely because in many cases no list of the reference population is available.

**Advantages**

 It reduces the field cost.

 Inferential statistics may be used.

 Sample may be comprehensive and representative of population.

 Observations of the sample may be used for drawing conclusions and ngeneralizations.

**Disadvantages**

This is not free from error, since there is subjectivity due to different ways of systematic list by different individuals. Knowledge of population is essential. This method can’t ensure the representativeness.

1. **Stratified Sampling**

It is an improvement over the earlier method. When employing this technique, the researcher divides his population in strata on the basis of some characteristics and from each of these smaller homogeneous groups (strata) draws at random a predetermined number of units. Researcher should choose that characteristic or criterion which seems to be more relevant in his research work.

As already mentioned with regard to sampling error, the accuracy of sampling estimates depends – among other things – on sample size and on the degree of variability in the distribution of the phenomenon studied within the reference population. Accuracy declines as sample size diminishes and as the distribution variability of the phenomenon increases. This means that if the variability of the phenomenon under investigation is very high, then the sample analyzed will need to be larger, in order to maintain a certain level of accuracy in the estimate. Alternatively, if the phenomenon displays areas of greater homogeneity, it is possible to increase sample *efficiency* (greater accuracy for the same sample size) by adopting *stratified sampling*.

This sampling design is organized in three phases. First of all, the reference population is subdivided into sub-populations (called *strata*) that are as homogeneous as possible in terms of the phenomenon to be studied; this is done by using as a stratification criterion a variable that is correlated with the phenomenon. Second, a sample is selected from each stratum by means of a random procedure. Finally, the samples drawn from each stratum are pooled in order to produce an overall sample. Within each single stratum, subjects are selected by means of a simple random sampling procedure.

**Advantages**

o It is (more precisely) a good representative of the population.

o It is an objective method of sampling.

o Observations can be used for inferential purpose.

**Disadvantages**

o Serious disadvantage of this method is that it is difficult for the researcher to decide the relevant criterion for stratification.

o Only one criterion can be used for stratification, but it generally seems more than one criterion relevant for stratification.

o It is costly and time consuming method.

o Selected sample may be representative with reference to the used criterion but not for the other.

**4. *Multistage sampling***

This technique does not offer greater efficiency than simple random sampling, but it does simplify the selection procedure and reduce the cost of data-collection. Multistage sampling is the only viable option in some situations, such as when a complete list of the reference population is unavailable, or when – on account of the extreme dispersion of the population – the members of the sample produced by simple random sampling or stratified sampling would be spread over too vast an area (e.g. the whole country) and thus difficult to reach.

In multistage sampling, the population is subdivided into hierarchical levels, which are selected successively through a process of ‘narrowing down’. For example, in order to draw up a national sample of elementary school teachers, we would first subdivide this population into *primary units*, constituted by the education districts to which the teachers belong, and then into *secondary units*, constituted by the teachers themselves.

Multistage sampling offers two main advantages: (a) there is no need to have a list of the entire reference population, but only lists of the sub-populations of the higher order units selected; and (b) data collection involves only the units selected, thereby reducing costs considerably. There are, however, disadvantages. Multistage sampling leads to a loss of efficiency, in that cases belonging to a single higher-order unit tend partly to resemble one another. Moreover, the statistical theory involved in this sampling design is somewhat complex.

***5. Cluster Sampling***

This procedure is used when the population can be subdivided into groups. This is the case of families, school classes, works departments, hospital wards and so on. Such groupings are called clusters, hence the name *cluster sampling*. In this procedure, it is not the units (the individuals) that are selected, but the clusters; all the units belonging to the clusters selected are then included in the sample. The procedure is extremely useful when – as in this example – no list of the units is available, but clusters can be selected by means of a random procedure.

**Sample Size**

The general rule for samples is the bigger the better. Here are three important considerations.

 First, decisions about sample sizes depend on how widely dispersed are the population characteristics. In our example, if all members of the population were the same age, we would need a sample of only one to estimate the mean age of the population. As the age distribution becomes wider, so the sample size must be increased, that is, if age is a critical variable.

 Second, we need to know what the risks are in making an incorrect estimation of a population parameter. In other words, how much sampling error can be tolerated? If the consequences of being wrong are serious, the sample size will need to be larger than where they are not as serious. For example, making an inaccurate estimate of the mean age of a population may not be as serious as determining whether a drug is safe to use.

 Third, the various methods of data analysis have different requirements as far as the minimum number is concerned. This is particularly critical when nominal level data are analyzed using cross-tabulations. This consideration is independent of the accuracy of the estimate of population characteristics. The general rule is that nominal-level data require larger numbers than or d in all level data and certainly larger than the methods appropriate for interval-level and ratio-level data. A rule of thumb for nominal-level data is that the cells of a cross-tabulation need an average of 10. Hence, the sample size can be determined by the combination of the number of categories in two variables. Population size, budget and time are also another factor determining the sample size of the study.

**Strategies for Determining Size of Respondents**

In fact, the sample size depends on several factors: purpose of the study, the type of population (varied or identical), availability of equipment and technical people, n recourses allotted to the study in terms of time and money, and the level of precession required. Here we should bear in mind that determining sample size is a very important issue because samples that are too large may waste time, resources and money, while samples that are too small may lead to inaccurate results.

There are several approaches to determining the size of the respondents. These include using a census for small populations, imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size. Each strategy is discussed below briefly.

***Using a Census for Small Populations***

One approach is a census survey in which we consider the entire population as respondent. It is very difficult and even sometimes impossible to carryout census survey for large population as it is time and money consuming. However, for a small population *(eg. 200 or less),* census yields more precise research output than sampling does as it touches each and every population under study. Census survey also eliminates sampling error and provides data on all the individuals in the population.

***Using a Sample Size of a Similar Study***

Another approach is to use the same sample size as those of other methodologically very sounding studies similar to the one you plan to carry out. In this case you must be very careful that you may run the risk of repeating errors that were made in determining the sample size for another study if you fail to deeply review the procedures employed in the earlier studies. However, a review of the literature in your discipline can provide guidance about typical sample sizes which are used.

***Using Formulas to Calculate a Sample Size***

The minimum sample size required for a very large population (N>10,000) is n = 􀭞􀬶􀭮(􀬵􀬿􀭮) 􀭛􀬶 where Z=1.96 is the values of Z tabulated at a 95% confidence interval, p= 50% or 0.5 where there is no information about the estimate of the proportion to be studied, W=+0.05 as the margin of error of the estimate proportion. However, when the population is less than 10,000 the sample size from the above formula used with some adjustment, the adjusted sample size n= ( 􀭬 􀬵􀬾􁉀􀱤 􀱊􁉁 ).

**Response Rate**

At the beginning of the chapter I raised the issue of the need to achieve a very high *response rate* before it is appropriate to use inferential analysis, 85 per cent in fact. The reason for insisting on such a high figure is that as the response rate declines, the possibility of a sample becoming unreliable, or being even more biased, increases. If non respondents or non-contacts are not typical of the sample as a whole, their absence will change the characteristics of the remaining sample. For example, if non respondents tend to be elderly people, then the age distribution of the sample will be distorted and its ability to accurately estimate the mean age of the population will be jeopardized. Non- response resulted a significant bias when the following two situation are fulfilled.

1. When non- respondents differ significantly from respondents

2. When non- respondents constitute a significant proportion of the sample. The issue of non-response should be considered during the planning stage of the study i.e. below 15%.

*Maintaining non-response at a low level:*

1. Training data collectors to contact respondents in respectful way, convince the importance of a given study

2. Offering incentive

3. Repeated attempt

4. Recording the number of non-respondents

**Types of Errors in Sampling**

While taking a sample, the result will not exactly equal the correct results for the whole population.

**1. Sampling Error (Random Error)**

Created being sample is unable to represent the whole characteristics of the population. This type of error arises from the sampling process itself and reduced by increasing the size of the sample.

**2. Non-sampling Error (Bias)**

It is systematic error in the design of a sampling procedure. Reduced or eliminated by careful design of the sampling procedure than by increasing the sample size. Possible source of bias in sampling includes accessibility bias, voluntary bias and non-response.

**Chapter Four: Instrument Design: Questionnaire Construction**

**4.1 The questionnaire**

The questionnaire is a measuring instrument. It measures someone’s behavior or attitude.

**1. Question Text**

The question text is the most important part of the question. This is what the respondents respond to. If they do not understand the question, they will not give the right answer, or they will give no answer at all. Rules considered in constructing questionnaire text:

1. Use Familiar Wording. The question text must use words that are familiar to those who have to answer them. Particularly, questionnaire designers must be careful not to use jargon that is familiar to themselves but not to the respondents. Indefinite words like “usually,” “regularly,” “frequently,” “often,” “recently,” and “rarely” must be avoided if there is no additional text explaining what they mean. How regular is regularly? How frequent is frequently? These words do not have the same meaning for every respondent. One respondent may interpret “regularly” as every day, while it could mean once a month to another respondent.

2. Avoid Ambiguous Questions. If the question text is such that different respondents may interpret the question differently, their answers will not be comparable. For example, if a question asks about income, it must be clear whether it is about weekly, monthly, or annual income. It must also be clear whether the respondents should specify their income before or after tax has been deducted. Vague wording may also lead to interpretation problems.

3. Avoid Long Question Texts. A respondent attempting to comprehend a long question may leave out part of the text and thus change the meaning of the question. Long texts may also cause respondent fatigue, resulting in a decreased motivation to continue. Of course, the question text should not be so short that it becomes ambiguous.

4. Avoid Recall Questions as much as Possible. Questions requiring recall of events that have happened in the past are a source of errors.

5. Avoid Leading Questions. A leading question is a question that is not asked in a neutral way but leads the respondents in the direction of a specific answer. For example, the question contains a reference to the “majority of people.” It suggests that it is socially undesirable to not agree. Do you agree with the majority of people that the quality of the health care in the country is falling? A question can also become leading by including the opinion of experts in questions text, such as most doctors say that cigarette smoking cause lung cancer. Do you agree? Such information should be objective and neutral and should not influence respondents in a specific direction.

6. Avoid Sensitive Questions. Sensitive questions should be avoided as much as possible. Respondents may refuse to provide information on topics such as income or health. Respondents may also avoid giving an answer that is socially undesirable.

7. Avoid Double Questions (or Double-Barreled Questions). A question must ask one thing at a time. For example, the question Do you think that people should eat less and exercise more? Actually consists of two questions: “Do you think that people should eat less?” and “Do you think that people should exercise more?”

8. Avoid Hypothetical Questions. It is difficult for people to answer questions about imaginary situations, as they relate to circumstances they have never experienced. At best, the answer is guesswork and a total lie at worst. Here is an example of a hypothetical question: If you were the president of the country, how would you stop crime?

**2. Answer Types**

Another important aspect of a survey question is the way in which the question must be answered. An open question is a simple question to ask. It allows respondents to answer the question completely in their own words. An open question is typically used in situations where respondents should be able to express themselves freely. Asking an open question may also lead to vague answers. Processing the answers to open questions is cumbersome, particularly if the answer is written down on a paper form.

Entering such answers in the computer takes effort, and even more if the written text is not very well readable. Furthermore, analyzing answers to open questions is not very straightforward. It is often done manually because there is no intelligent software that can do this automatically.

Considering the potential problems mentioned above, open questions should be avoided wherever possible. However, there are situations where there is no alternative. An example is a question asking for the occupation of the respondent. A list containing all possible occupations would be very long. It could easily have thousands of entries. A closed question is used to measure qualitative variables. There is a list of possible answers corresponding to the categories of the variable. Respondents have to pick one possibility from the list. Of course, this requires the list to contain all possible answers:

|  |
| --- |
| **What is your present marital status?**  Never married……………………………….....……….1  Married…………………………………….……………2  Divorced………………………………………,..….……3  Separated……………………………………........……..4  Widowed…………………………………………….…..5 |

There will be problem if respondents cannot find their answer. One way to avoid such a problem is to add a category “other,” possibly also offering the option to enter the answer. Only one answer is allowed for a closed question. Respondents seek the easiest way to answer a question by simply selecting they don’t know option. Such behavior can be avoided by training interviewers to assist respondents to give a real answer and to avoid don’t know as much as possible. The option is not explicitly offered but is implicitly available. Only if respondents indicate that they really do not know the answer, the interviewer records this response as don’t know. All answer options have to be mutually exclusive and exhaustive. So respondents can always find one and only one option referring to their situation. Sometimes, however, there are closed questions in which respondents must have the possibility to select more than one option.

**2.1. Characteristics of a Good Questionnaire**

The following are the characteristics of a good questionnaire:

1. The covering letter of the questionnaire is drafted in a befriending tone and indicates its importance to the respondents.

2. The questionnaire contains directions which are clear and complete. Important items are clearly defined and each question deals with a single idea defined in unambiguous terms.

3. It is reasonable short, through comprehensive enough to secure all relevant information.

4. It is attractive in appearance, neatly arranged, clearly duplicated and free from typographical errors.

5. It avoids annoying or embracing questions, which arouse hostility in the respondent.

6. Items are arranged in categories which ensure easy and accurate responses.

7. Questions do not contain leading suggestions for the respondents and are objective in nature.

8. They are arranged in good order. Simple and general questions should precede the specific and complex ones. Questions that create favorable atmosphere should precede those that are personal and touch delicate points.

**3. Question Order**

Once all questions have been defined, they have to be included in the questionnaire in the proper order. The first aspect is grouping of questions. It is advised to keep questions about the same topic close together. The second aspect is the potential learning effect. An issue addressed early in the questionnaire may make respondents think about it.

Ideally, the early questions in a survey should be easy and pleasant to answer. Such questions encourage respondents to continue the survey. Whenever possible, difficult or sensitive questions should be asked near the end of the questionnaire. If these questions cause respondents to quit, at least many other questions have been answered.

**4. Questionnaire Testing**

Before a questionnaire can be used to collect data, it must be tested. Errors in the questionnaire may cause wrong questions to be asked and right questions to be skipped. Also, errors in the questions themselves may lead to errors in answers. This has the advantage that the interviewers can ask the respondents whether they have understood the questions, whether things were unclear to them, and why they gave specific answers.

A number of aspects of a questionnaire should be tested. May be the most important aspect is the validity of the question. Does the question measure what the researcher wants to measure? It is not simple to establish question validity in practice. A first step may be to determine the meaning of the question. It is important that the researcher and the respondent interpret the question exactly in the same way.

Another aspect of questionnaire testing is to check whether questions offer sufficient variation in answer possibilities. A survey question is not very interesting for analysis purposes if all respondents give the same answer.

**5. Non-response**

Nonresponsive occurs when elements in the selected sample that are also eligible for the survey do not provide the requested information or that the provided information is not usable.

**MODES OF SURVEY ADMINISTRATION**

􀀀􀀀Personal (Face-to-Face)

􀀀􀀀Telephone

􀀀􀀀Mail

􀀀􀀀Web

􀀀􀀀Combination of Methods

**PERSONAL INTERVIEWING**

***Advantages***:

􀀀􀀀Generally yields highest cooperation and lowest refusal rates

􀀀􀀀Allows for longer, more complex interviews

􀀀􀀀High response quality

􀀀􀀀Takes advantage of interviewer presence

􀀀􀀀Multi-method data collection

***Disadvantages***:

􀀀􀀀Most costly mode of administration

􀀀􀀀Longer data collection period

􀀀􀀀Interviewer concerns

**TELEPHONE INTERVIEWING**

***Advantages:***

􀀀􀀀Less expensive than personal interviews

􀀀􀀀Shorter data collection period than personal interviews

􀀀􀀀Better control and supervision of interviewers (vs. personal)

􀀀􀀀Better response rate than mail for list samples

***Disadvantages:***

􀀀􀀀Biased against households without telephones, unlisted numbers

􀀀􀀀Non response

􀀀􀀀Difficult to administer questionnaires on sensitive or complex topics

**MAIL SURVEYS**

***Advantages***:

􀀀􀀀Generally lowest cost

􀀀􀀀Can be administered by smaller team of people (no field staff)

􀀀􀀀Access to otherwise difficult to locate, busy populations

􀀀􀀀Respondents can look up information or consult with others

***Disadvantages:***

􀀀􀀀Most difficult to obtain cooperation

􀀀􀀀No interviewer involved in collection of data

􀀀􀀀Need good sample

􀀀􀀀More likely to need an incentive for respondents

􀀀􀀀Slower data collection period than telephone

Inferential Analysis: From Sample to Population

**Chapter Five: Statistical Analysis of Data**

Most commonly used statistical data analysis methods are:

 Calculating frequency distribution usually in percentages of items under study.

 Testing data for normality of distribution skewness and kurtosis.

 Calculating percentiles and percentile ranks.

 Calculating measures of central tendency-mean, median and mode and establishing norms.

 Calculating measures of dispersion-standard deviation, mean deviation, quartile N deviation and range.

 Calculating measures of relationship-coefficient of correlation.

 Graphical presentation of data-Frequency polygon curve, Histogram, Cumulative frequency polygon and O give etc.

**5.1. Types of Analysis**

Various methods of data analysis are used to describe the characteristics of social phenomena, and to understand, explain and predict patterns in social life or in the relationships between aspects of social phenomena. Hence, analysis can be divided into four types: univariate descriptive, bivariate descriptive, explanatory and inferential.

**5.1.1. Univariate Descriptive**

Univariate descriptive statistics include:

– Percentages, averages, and various charts and graphs.

– *Univariate descriptive analysis* is used to represent the characteristics of some social phenomenon (e.g. student academic performance on a particular course).

This can be done in a number of ways:

o by counting the frequency with which some characteristic occurs (e.g. the total marks of students receive on a particular course);

o by grouping scores of a certain range into categories and presenting these frequencies in pictorial or graphical form (e.g. student’s total marks);

o by calculating measures of central tendency (e.g. the mean marks obtained by students on the course); and

o by graphing and/or calculating the spread of frequencies around this centre point (e.g. plotting a line graph of the frequency with which particular marks were obtained, or calculating a statistic that measures the dispersion around the mean).

***1. Frequency Distribution***

Tabulation is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious.

In statistics, a **frequency distribution** is a list of values that a variable takes in a sample. It is usually a list, ordered by quantity, showing the number of times each value appears. To establish frequencies of occurrence, data must be in categories. Frequency counts summarize data that have been collected in *nominal* categories, *ordinal* categories, in whole numbers (*discrete* data), and in *continuous* values or scores that have been grouped into categories.

**2. Graphic presentation**

Several types of statistical/data presentation tools (graphic aids) exist. These include bar graphs, histograms, scattered diagrams, pie-chart, line graphs and tables. In addition to being presented in tabular form, frequency counts and distributions can be presented pictorially or graphically. The commonly used methods are bar charts and pie charts for categorical variables, and histograms and line graphs for metric variables consisting of discrete or grouped data. The shape of the graph offers easy and immediate answers to several questions such as the variations of the data distribution and trends. Graphic presentation, therefore, serves as an easy technique for quick and effective comparison b/n two or more variables.

Generally, presentation of data in diagrams has the following advantages:

1. Diagrams give an attractive and elegant presentation of data

2. Diagrams leave good visual impact and facilitate comparisons

3. Interpretations from diagrams save time

4. Diagrams simplify complexity and easily depict the characteristics of the data *Bar charts* are used to represent the frequency counts of nominal-level and ordinal-level data, or grouped data from interval-level or ratio-level variables.

*Pie charts* provide an alternative way of presenting frequency counts pictorially. The bars of a chart are replaced by segments of a circle such that the area of a segment corresponds to the frequency in that category. They are alternative presentations to the bar charts. Because of the continuous nature of their form of measurement, metric variables using discrete or grouped data can be represented pictorially either as histograms or as continuous lines. The *histogram* is a variation on the bar chart such that the bars touch each other to create a stair pattern. Instead of having gaps between the bars, they are any histogram with categories of equal width or interval can be converted into a *line graph*, also known as an area graph or a frequency polygon, by joining up the midpoints of the bars in a histogram with straight lines.

**3. Measures of Central Tendency and Variation or Dispersion**

In the previous sections, we discussed how raw data can be collected, organized and presented in terms of tables, charts and frequency distribution in order to be easily understood and analyzed. Although frequency distribution and corresponding graphical presentations make raw data more meaningful, yet they fail to identify major properties that describe a set of quantitative data.

These are:

1. Central value of a set of data called c***entral tendency***

2. The extent to which numerical values are dispersed around the central value, called **variation or dispersion** in terms of single distances of individual observation from central values.

A. C***entral Tendency***

***1. Mean***

Mean is a central value which is computed by taking into consideration all the observations or all recorded values. Mathematically, mean of a list of numerical observations is the sum of the entire observation divided by the number of items in the list.

***Merits and Demerits of Arithmetic Mean***

***Merits***

o The calculation of arithmetic mean is simple and it is unique,

o It is clear and unambiguous since every data set has one and only one mean value

o The calculation of arithmetic mean is based on all values given in the data set

o The arithmetic mean is reliable single value that reflects all values in the data set.

o The arithmetic mean is least affected by fluctuations in sample size. Its value determined from various samples drawn from a population vary by the least possible amount

o It used for more rigorous further statistical analysis

o Arithmetic mean is a stable average

**Demerits**

 The value of arithmetic mean cannot be calculated accurately for open-ended class intervals

 It is affected by the extreme values which are not the exact representative of the data set.

 For large data sets the calculations of arithmetic mean may sometimes be difficult and tedious as every element is used in the calculation

 It cannot be calculated for qualitative characteristics such as intelligence, beauty and loyalty

 Arithmetic mean cannot be determined by inspection

B. **Median**

**Median** is defined as a middle value in an ordered sequence of data in the sense that half of the observations are smaller and half are larger than this value. If the number of observations (n) is odd number, then the median (Med) is represented by the numerical value corresponding to the positioning point of the *order observation* If the number of observations (n) is an even number, then the median is defined as the arithmetic mean of the numerical values of the *observations in the data array*

***Merits***

 *Easy to* calculate and understand even for professionals with low level of mathematics and statistics

 It is not affected by the extreme values in a data set

 Median can be computed while dealing with a distribution with open ended classes

 Median can sometimes be located by simple inspection

**Demerits**

o In case of even number of observations for ungrouped data, median cannot be determined exactly

o Median, being a positioning value, is not based on each item in a data set

o Median is not suitable for further mathematical treatment

o Median is more affected by fluctuations of sampling as compared to arithmetic mean.

C. **Mode**

In statistics, the **mode** is the value that occurs most frequently in a data set or a probability distribution. Though a poor measure of central tendency, mode has some advantages. It can be calculated only by inspection from a simple frequency distribution. It is also unaffected by the presence of extreme values of a data set and can be calculated from frequency distribution with open ended classes. Mode has little or no use for farther statistical analysis.

**Merits**

1. Mode value is easy to understand and to calculate

2. Mode class can be inspected by inspection

3. Mode is not affected by extreme values in the distribution

4. Mode value can be calculated for open-end frequency distributions

**Demerits**

1. A data set may have more than one mode value which makes the comparison and interpretation more difficult

1. It is difficult to locate modal class in the case of multi-modal frequency distributions

2. Mode is not used for further rigorous statistical analysis

**2. Measures of Variability**

Measure of central tendency-mean, median and mode give us merely an idea of the general achievement of the group as a whole, and does not show how the individual scores are spread out. Therefore, through measures of central tendency we are unable to know much about the distribution of scores in a series or characteristics on items in a group.

Test Scores of Group A (boys) – 40, 38, 36, 17, 20, 19, 18, 3, 5, 4

Test Scores of Group B (girls) – 19, 20, 22, 18, 21, 23, 17, 20, 22, 18.

Now the value of the Mean in both the above cases is 20 and thus, so far as the mean goes, there is no difference in the performance of the two groups. Now the question arises, can we take both sets of scores as identical? Definitely there is a lot of difference between the performances of two groups. Whereas the test scores of group A are found to range from 4 to 40, the scores in group B range from 18 to 23. Individuals in this latter group are less variable than those of the former.

**Different Measures of Variability or Dispersion**

These are chiefly, four measures of indicating variability or dispersion within the set of scores. They are:

(*a*) The Range (R)

(*b*) The Quartile Deviation (Q)

(*c*) The Standard Deviation (SD)

Each of the above measures of variability tells us how the individual scores are scattered or spread over throughout the distribution or gives data.

**A. Range (R)**

Range is the simplest measure of variability or dispersion. It is calculated by subtracting the lowest scores in the series from the highest. It takes only extreme scores into consideration and tells nothing about the variation of the individual items.

**B. Quartile Deviation (Q)**

It is computed by the formula *Q* = (*Q*3 – *Q*1), where *Q*1 and *Q*3 represent the 1st and 3rd quartiles of the distribution under consideration. The amount *Q*3 and *Q*1 is nothing but the difference of range between 3rd and 1st quartile. It is designated as the inter quartile range.

***C. Standard Deviation***

Standard deviation, ***also known as root mean squared deviation***, explains the average amount of variation on either side of the mean. It is considered to be the best measure of dispersion and is used most widely even in advanced techniques. The population ( ) and the sample (s) standard deviations are the positive square roots of their respective variances and have the desirable property of being in the same units as the data.

**5.1.2. Bivariate Descriptive Analysis**

This session explores the major ways in which relationships between various types of variables can be demonstrated and their strength measured. Distributions tell us about the characteristics of single variables, while measures of association tell us about the connections between variables. However, they do not tell us anything about how one variable might influence another variable. Two variables are said to be associated if the values of one variable vary or change together with the values of the other variable. For example, persons with a high level of education may also have a high income, and those with a low education a low income. As level of education increases, so does level of income. The relationship between two variables can take different forms. Sometimes there may be *no relationship*. For example, persons with a high level of education are no more likely than those with a low or moderate education to have a high income. When a high position or score on one variable *is* associated with a high position or score on the other variable, and vice versa, a *positive relationship* is said to exist. Alternatively, if a high position or score on one variable is associated with a low position or score on the other variable, and vice versa, then there is a *negative relationship*.

**5.1.3. Explanatory Analysis: Looking for Influences**

The establishment of associations between variables is an important part of descriptive analysis. To go beyond describing characteristics and establishing relationships, that is, to go beyond answering ‘what’ questions to addressing ‘why’ questions, takes us into the complex and difficult territory of *explanatory analysis* and the much disputed notion of *causation*. The existence of an association is a necessary but not a sufficient condition for explanatory analysis. Association on its own does not allow us to infer that one variable has an influence on the other. It is a common belief that establishing an explanation involves finding the cause or causes for the patterns and sequences in social life.

One variable, the values of which are to be explained, is referred to as the *dependent* or *outcome variable*, while those that are involved in producing these values are referred to as *independent* or *predictor variables*. Stated differently, the values of the dependent variable are influenced or predicted by the values of the independent variable or variables.

**5.1.4. Inferential Statistics**

A statistic computed from a sample may be used to estimate a parameter, the corresponding value in the population which it is selected. Inferential statistics is when we make inferences to hypothesis testing, determine relationships, and make predictions out of the analyzed data. *Inferential analysis* is used with data obtained from a sample to estimate the characteristics of or patterns in the population from which the sample was drawn. This kind of analysis is only appropriate when the sample is drawn using probability or random selection procedures.

The sample data are usually just presented and assumed to be the same in the population. In contrast, a great deal of attention is given to testing whether relationships found in a sample can be expected to exist in a population. This is done by using *tests of significance*.

***5.2. Correlation and Regression***

For both regression and correlation studies, the number of variables may be two (Bivariate analysis) or more (Multivariate analysis).

***5.2.1. Correlation Analysis***

In statistics, **correlation** indicates the strength and direction of the mutual interdependence of two or more variables. The relative strength***s of*** relationships are identified by a measure referred to as ***Correlation Coefficient or Coefficient of*** ***Correlation***. The symbol r **stands** for the coefficient of correlation. It is always between -1.0 and +1.0. If the coefficient of correlation is negative, we have an inverse relationship; if it's direct, the relationship is positive. A correlation coefficient of -1.00 tells you that there is a *perfect negative relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *decrease* in values on the other variable.

In other words, as one variable goes up, the other goes in the opposite direction (it goes down). A correlation coefficient of +1.00 tells you that there is a *perfect positive* *relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *increase* in values on the other variable. In other words, as one variable goes up so does the other. A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can’t really say anything about what happens to the other variable.

One way to get a general idea about whether or not two variables are related is to plot them on a “scatter plot”. If the dots on the scatter plot tend to go from the lower left to the upper right it means that as one variable goes up the other variable tends to go up also. This is a called a “*positive relationship*”.

Generally, through correlation analysis we can found the following information:

a) Does a relationship exist?

b) If so, is it a positive or a negative relationship? and

c) Is it a strong or a weak relationship?

***5.2.2. Regression Analysis***

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships.

1. ***Simple Regression***

Regression analysis with a single explanatory variable is termed “simple regression.”In reality, any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed “omitted variables bias”). Thus, the tentative hypothesis is that higher levels of education cause higher levels of earnings, other things being equal.

**B. *Multiple Regression Model***

The general purpose of multiple regressions (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

Plainly, earnings are affected by a variety of factors in addition to years of schooling, factors that were aggregated into the noise term in the simple regression model above. “Multiple regressions” is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables. When Should I Use Multiple Regressions?

1. The criterion variable that you are seeking to predict should be measured on a continuous scale (such as interval or ratio scale). There is a separate regression method called logistic regression that can be used for dichotomous dependent variables (not covered here).

2. The predictor variables that you select should be measured on a ratio, interval, or ordinal scale. A nominal predictor variable is legitimate but only if it is dichotomous, i.e. there are no more than two categories. For example, sex is acceptable (where male is coded as 1 and female as 0) but gender identity (masculine, feminine and androgynous) could not be coded as a single variable. Instead, you would create three different variables each with two categories (masculine/not masculine; feminine/not feminine and androgynous/not androgynous). The term dummy variable is used to describe this type of dichotomous variable.

3. Multiple regression requires a large number of observations. The number of cases (participants) must substantially exceed the number of predictor variables you are using in your regression. The absolute minimum is that you have five times as many participants as predictor variables. A more acceptable ratio is 10:1, but some people argue that this should be as high as 40:1 for some statistical selection methods.

**5.3. Hypothesis Testing**

Generally speaking, there are two types of tests conducted when using inferential procedures:

 tests of group differences

 tests of association.

With a *test of group differences*, you want to know whether two populations differ with respect to their mean scores on some response variable. The present experiment leads to a test of group differences because you want to know whether the average amount of insurance sold in the population of difficult-goal agents is different from the average amount sold in the population of easy-goal agents. In this case, two or more distinct populations are being compared with respect to their mean scores on a single response variable.

With a *test of association*, there is a single population of individuals and you want to know whether there is a relationship between two or more variables within this population. Perhaps the best-known test of association involves testing the significance of a correlation coefficient.

**5.3.1. Types of Hypotheses**

Two different types of hypotheses are relevant to most statistical tests. The first is called the *null hypothesis*, which is often abbreviated as H0. The null hypothesis is a statement that, in the population(s) being studied, there are either (a) no differences between the group means, or (b) no relationships between the measured variables. With a test of group differences, the null hypothesis states that, in the population, there are no differences between any of the groups being studied with respect to their mean scores on the response variable. In the experiment in which a difficult-goal condition is being compared to an easy-goal condition, the following null hypothesis might be used: H0: In the population, the amount of insurance sold by individuals assigned difficult goals does not differ from the amount of insurance sold by individuals assigned easy goals. This null hypothesis can also be expressed with symbols in the following way: H0: M1 = M2

In contrast to the null hypothesis, there is also an *alternative hypothesis* (H1) that states the opposite of the null. The alternative hypothesis is a statement that there is a difference between the means, or that there is a relationship between the variables, in the population(s) being studied. Perhaps the most common alternative hypothesis is a *non directional alternative hypothesis*. With a test of group differences, a non directional alternative hypothesis predicts that the means for the various populations differ, but makes no specific prediction as to which mean will be relatively high and which will be relatively low. H1: In the population, individuals assigned difficult goals differ from individuals assigned easy goals with respect to the mean amount of insurance sold. This alternative hypothesis can also be expressed with symbols in the following way: H1: M1 ≠not equal to M2. In contrast, a *directional alternative hypothesis* makes a more specific prediction regarding the expected outcome of the analysis. With a test of group differences, a directional alternative hypothesis not only predicts that the population means differ, but also predicts which population means will be relatively high and which will be relatively low. H1: The average amount of insurance sold is higher in the population of individuals assigned difficult goals than in the population of individuals assigned easy goals. This hypothesis can be symbolically represented as follows: H1: M1 > M2.

There is an important advantage associated with the use of directional alternative hypotheses compared to non directional hypotheses. Directional hypotheses allow researchers to perform *one-sided* statistical tests (also called one-tailed tests), which are relatively powerful. However, directional hypotheses should be stated only when they can be justified on the basis of theory, prior research, or some other grounds.

**The *p* Value**

Hypothesis testing is a process to determine whether you can reject a null hypothesis with an acceptable level of confidence. When analyzing data with JMP, you look at the results for two pieces of information that are critical for this purpose:

1. the obtained (calculated) statistic

2. the probability (*p*) value associated with that statistic.

The *p* value indicates the probability that you would obtain the present results if the null hypothesis were true*.* If the *p* value is very small, you reject the null hypothesis. In rejecting the null hypothesis, you have tentatively accepted the alternative hypothesis. A *p* value of 0.05 is one of the most commonly accepted cutoff values. Typically, when researchers obtain a *p* value *larger* than 0.05 (such as 0.13 or 0.37), they fail to reject the null hypothesis. The 0.05 level of significance is not an absolute rule that must be followed in all cases, but it is serviceable for most types of investigations likely to be conducted in the social sciences and other areas.

***5.3.2. Procedure for Hypothesis Testing***

This refers to the steps required to test the validity of the claim or assumption about the sample statistic. Hence, the general procedure for any hypothesis testing is summarized below.

***Step 1: State the null hypothesis (Ho) and alternative hypothesis (H1)***

Theoretically, hypothesis testing requires that the null hypothesis to be considered true or no difference until it is proved false on the basis of results observed from the sample data. The null hypothesis is always expressed in the form of an equation making claim regarding the specific value of the population parameter.

***Step 2: State the Level of Significance or (alpha) for the Test***

The level of significance is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is specified in terms of the level of probability of the null hypothesis being wrong or rejected.

***Step 3: Establish Critical or Rejection Region***

***Step 4: Calculate the Suitable Test Statistic***

The value of the test statistic is calculated from the distribution of sample statistic.

***Step 5: Reach a Conclusion***

Compare the calculated value of the test statistic with the critical value (also called standard table value or tabulated value). The decision rules for null hypothesis are as follows:

|Value Cal |>Value Table; Reject the *Ho*

|Value Cal < |Value Table; Accept the *Ho*

**One-tailed and two-tailed Tests**

There are two types of tests referred to as the one-tailed and two-tailed tests. The type of tests depends on the way the hypotheses are formulated.

***a. Two-tailed test is when null and alternative hypotheses are stated as:***

Ho: μ = μo and H1: μ not equal to μo. This implies that any deviation (either on the lower or higher side) of the calculated value of test statistic from the hypothesized value leads to rejection of the null hypothesis, Ho.

***b. One-tailed test is when null and alternative hypotheses are stated as:***

*Ho: μ <μo and H1: μ >μo (Right-tailed test), or*

*Ho: μ> μo and H1: μ < μo (Left-tailed test),*

This implies that the value of sample statistic is either higher or lower than the hypothesized parameter value. This leads to the rejection of null hypothesis for significant deviation from the specified value in one direction or tail of the curve of sampling distribution.

***5.3.3. Common Types of Hypothesis Testing***

***A. Chi-square (2) Test***

This is also one of the non-parametric or distribution free statistical tests which go back to 1900, when Karl Pearson used it for frequency data classified into k-mutually exclusive categories. It is usually represented by**,** a Greek letter **Chi**. The test of significance of correlation coefficient means to test the hypothesis, whether or not the correlation coefficient is zero in population. This means we test: *Ho*: p = 0 (There is no correlation) vs *H1*: p 0 (There exist a correlation)

1. ***Independent T-test***

Is applicable in understanding the mean variation between two independent groups. Equal variance assumed when Levene’s, p > 0.05. If Levene’s p<0.05, equal variance was not assumed. Independent t-test is employed to testify the null hypothesis that states there is no mean score variation between two independent groups.

***C. Analysis of Variance (ANOVA)***

Analysis of variance (ANOVA), developed by Sir Ronald Fisher, helps to test the differences between three or more sample means drawn from two or more sets of populations. Here we test null hypothesis *(Ho)* that three or more sets of population or populations from which samples are drawn have equal (homogeneous) means against the alternative hypothesis *(H1)* that population means are not equal at all.

H0: μ = μ2 = μk

H1: Not all μj are equal (j = 1, 2, 3, k)

Hence, the null and alternative hypothesis of population means imply that the null hypothesis should be rejected if any of the samples means is different from others.

The assumptions for the analysis of variance are:

A. Each population is normal distribution

B. The sets of populations from which the samples are drawn have equal variances

C. Each sample is drawn randomly and is independent of other samples.

End