University of East London

Pioneering Futures Since 1898

UNICAF

UEL-DS-7006 Quantitative Data Analysis (QDA)

WEEK 2-Reading material

Sources of Data and Official Statistics; Handling Large Data Sets

Table of Contents

University of East London

Pioneering Futures Since 1898

UNICAF

UEL-DS-7006 Quantitative Data Analysis (QDA)

WEEK 2-Reading material

Sources of Data and Official Statistics; Handling Large Data Sets

Learning Objectives

By the end of this week, you will be able to;

- Understand how to collect data from different sources
- Be able to conduct interview method and its chief merit
- Perform data collection through questionnaire
- How to select appropriate method for data collection
- Differentiate between primary data and secondary data

INTRODUCTION

The task of data collection begins after a research problem has been defined and research design/ plan chalked out. While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz., primary and secondary. The primary data are those which are collected afresh and for the first time, and thus happen to be original in character (Daniel, 2019).

The secondary data, on the other hand, are those which have already been collected by someone else and which have already been passed through the statistical process. The researcher would have to decide which sort of data he would be using (thus collecting) for his study and accordingly he will have to select one or the other method of data collection. The methods of

collecting primary and secondary data differ since primary data are to be originally collected, while in case of secondary data the nature of data collection work is merely that of compilation. We describe the different methods of data collection, with the pros and cons of each method.

Collection of Primary Data

We collect primary data during doing experiments in an experimental research but in case we do research of the descriptive type and perform surveys, whether sample surveys or census surveys, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews. This, in other words, means that there are several methods of collecting primary data, particularly in surveys and descriptive researches. Important ones are (Tukey, 1977):

[1] observation method,

[2] interview method,

[3] through questionnaires,

[4] through schedules, and

[5] other methods which include

a. warranty cards;

b. distributor audits;

c. pantry audits;

d. consumer panels;

e. using mechanical devices;

f. through projective techniques;

g. depth interviews, and

h. content analysis.

The above methods have been explained briefly below separately

- **Observation Method**

The observation method is the most commonly used method especially in studies relating to behavioral sciences. Observation becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability. Under the observation method, the information is sought by way of investigator's own direct observation without asking from the respondent. For instance, in a study relating to consumer behavior, the investigator instead of asking the brand of wrist watch used by the respondent, may himself look at the watch. The main advantage of this method is that subjective bias is eliminated, if observation is done accurately (Kabir, 2016). Secondly, the information obtained under this method relates to what is currently happening; it is not complicated by either the past behavior or future intentions or attitudes. Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happens to be the case in the interview or the questionnaire method. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other However, observation method

has various limitations. Firstly, it is an expensive method. Secondly, the information provided by this method is very limited (Kabir, 2016). Thirdly, sometimes unforeseen factors may interfere with the observational task. At times, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

While using this method, the researcher should keep in mind things like: What should be observed? How the observations should be recorded? Or how the accuracy of observation can be ensured? In case the observation is characterized by a careful definition of the units to be observed, the style of recording the observed information, standardized conditions of observation and the selection of pertinent data of observation, then the observation is called as structured observation. But when observation is to take place without these characteristics to be thought of in advance, the same is termed as unstructured observation. Structured observation is considered appropriate in descriptive studies, whereas in an exploratory study the observational procedure is most likely to be relatively unstructured.

We often talk about participant and non-participant types of observation in the context of studies, particularly of social sciences. This distinction depends upon the observer's sharing or not sharing.

The life of the group he is observing. If the observer observes by making himself a member of the group, he is observing so that he can experience what the members of the group experience, the observation is called as the participant observation. But when the observer observes as a detached emissary without any attempt on his part to experience through participation what

others feel, the observation of this type is often termed as non-participant observation. (When the observer is observing in such a manner that his presence may be unknown to the people he is observing, such an observation is described as disguised observation).

There are several merits of the participant type of observation (Kabir, 2016):

(i) The researcher is enabled to record the natural behavior of the group.

(ii) (ii) The researcher can even gather information which could not easily be obtained if he observes in a disinterested fashion.

(iii) The researcher can even verify the truth of statements made by informants in the context of a questionnaire or a schedule. But there are also certain demerits of this type of observation viz., the observer may lose the objectivity to the extent he participates emotionally; the problem of observation control is not solved; and it may narrow-down the researcher's range of experience.

Sometimes we talk of controlled and uncontrolled observation. If the observation takes place in the natural setting, it may be termed as uncontrolled observation, but when observation takes place according to definite pre-arranged plans, involving experimental procedure, the same is then termed controlled observation. In non-controlled observation, no attempt is made to use precision instruments. The major aim of this type of observation is to get a spontaneous picture of life and persons. It tends to supply naturalness and completeness of behavior, allowing sufficient time for observing it. But in controlled observation, we use mechanical (or precision) instruments as aids to accuracy and standardization. Such observation tends to supply formalized

data upon which generalizations can be built with some degree of assurance. The main pitfall of non-controlled observation is that of subjective interpretation. There is also the danger of having the feeling that we know more about the observed phenomena than we do. Generally, controlled observation takes place in various experiments that are carried out in a laboratory or under controlled conditions, whereas uncontrolled observation is resorted to in case of exploratory researches.

- **Interview Method**

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews (Jafri, n.d).

(a) Personal interviews: Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons. This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation. In the case of direct personal investigation, the interviewer must collect the information personally from the sources concerned. He must be on the spot and must meet people from whom data have to be collected. Most of the commissions and committees appointed by government to carry on investigations make use of this method.

The method of collecting information through personal interviews is usually carried out in a structured way. As such we call the interviews as structured interviews. Such interviews involve the use of a set of predetermined questions and of highly standardized techniques of recording.

Thus, the interviewer in a structured interview follows a rigid procedure laid down, asking questions in a form and order prescribed. As against it, the unstructured interviews are characterized by a flexibility of approach to questioning. Unstructured interviews do not follow a system of pre-determined questions and standardized techniques of recording information. In a non-structured interview, the interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires. He may even change the sequence of questions. He has relatively greater freedom while recording the responses to include some aspects and exclude others. But this sort of flexibility results in lack of comparability of one interview with another and the analysis of unstructured responses becomes much more difficult and time consuming than that of the structured responses obtained in case of structured interviews (Kabir, 2016). Unstructured interviews also demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, however, happens to be the central technique of collecting information in case of exploratory or formularize research studies. But in case of descriptive studies, we quite often use the technique of structured interview because of its being more economical, providing a safe basis for generalization and requiring relatively lesser skill on the part of the interviewer.

We may as well talk about focused interview, clinical interview and the nondirective interview. Focused interview is meant to focus attention on the given experience of the respondent and its effects. The main task of the interviewer in case of a focused interview is to confine the respondent to a discussion of issues with which he seeks conversance. Such interviews are used

generally in the development of hypotheses and constitute a major type of unstructured interviews. The clinical interview is concerned with broad underlying feelings or motivations or with the course of individual's life experience. The method of eliciting information under it is generally left to the interviewer's discretion. In case of non-directive interview, the interviewer's function is simply to encourage the respondent to talk about the given topic with a bare minimum of direct questioning. Despite the variations in interview-techniques, the major advantages and weaknesses of personal interviews can be enumerated in a general way.

The chief merits of the interview method are as follows (Walker, 1940):

- More information and that too in greater depth can be obtained.
- Interviewer by his own skill can overcome the resistance, if any, of the respondents; the interview method can be made to yield an almost perfect sample of the general population.
- There is greater flexibility under this method as the opportunity to restructure questions is always there, especially in case of unstructured interviews.
- Observation method can as well be applied to recording verbal answers to various questions.
- Personal information can as well be obtained easily under this method.
- Samples can be controlled more effectively as there arises no difficulty of the missing returns; non-response generally remains very low.

- The interviewer can usually control which person(s) will answer the questions. This is not possible in mailed questionnaire approach. If so desired, group discussions may also be held.
- The interviewer may catch the informant off-guard and thus may secure the most spontaneous reactions than would be the case if mailed questionnaire is used.
- The language of the interview can be adopted to the ability or educational level of the person interviewed and as such misinterpretations concerning questions can be avoided.
- The interviewer can collect supplementary information about the respondent's personal characteristics and environment which is often of great value in interpreting results.

(b) Telephone interviews: This method of collecting information consists in contacting respondents on telephone itself. It is not a very widely used method, but plays important part in industrial surveys, particularly in developed regions due to the following merits It is faster than other methods i.e., a quick way of obtaining information.

1. It is cheaper than personal interviewing method; here the cost per response is relatively low.
2. Recall is easy; callbacks are simple and economical.
3. There is a higher rate of response than what we have in mailing method; the non-response is generally very low.
4. Replies can be recorded without causing embarrassment to respondents.
5. Interviewer can explain requirements more easily.

6. At times, access can be gained to respondents who otherwise cannot be contacted for one reason or the other.

7. No field staff is required.

8. Representative and wider distribution of sample is possible. But this system of collecting information is not free from demerits. Some of these may be highlighted.

1. Little time is given to respondents for considered answers; interview period is not likely to exceed five minutes in most cases.

2. Surveys are restricted to respondents who have telephone facilities.

3. Extensive geographical coverage may get restricted by cost considerations.

4. It is not suitable for intensive surveys where comprehensive answers are required to various questions.

5. Possibility of the bias of the interviewer is relatively more.

6. Questions must be short and to the point; probes are difficult to handle.

Collection of Data Through Questionnaire

This method of data collection is quite popular, particularly in case of big enquiries. It is being adopted by private individuals, research workers, private and public organizations and even by governments (Bohrnstedt & Knoke, 1982).

In this method a questionnaire is sent (usually by post) to the persons concerned with a request to answer the questions and return the questionnaire. A questionnaire consists of several questions

printed or typed in a definite order on a form or set of forms. The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents must answer the questions on their own. The method of collecting data by mailing the questionnaires to respondents is most extensively employed in various economic and business surveys. The merits claimed on behalf of this method are as follows:

i. There is low cost even when the universe is large and is widely spread geographically.

ii. It is free from the bias of the interviewer; answers are in respondents' own words.

iii. Respondents have adequate time to give well thought out answers.

iv. Respondents, who are not easily approachable, can also be reached conveniently.

v. Large samples can be made use of and thus the results can be made more dependable and reliable.

The main demerits of this system can also be listed here:

- Low rate of return of the duly filled in questionnaires; bias due to no response is often indeterminate.
- It can be used only when respondents are educated and cooperating.
- The control over questionnaire may be lost once it is sent.
- There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been dispatched.

- There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
- It is difficult to know whether willing respondents are truly representative.
- This method is likely to be the slowest of all.

Collection of Data Through Schedules

This method of data collection is very much like the collection of data through questionnaire, with little difference which lies in the fact that schedules are being filled in by the enumerators who are specially appointed for the purpose (Jenkins et al., 1975). These enumerators along with schedules go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma. In certain situations, schedules may be handed over to respondents and enumerators may help them in recording their answers to various questions in the said schedules. Enumerators explain the aims and objects of the investigation and remove the difficulties which any respondent may feel in understanding the implications of a question or the definition or concept of difficult terms. This method requires the selection of enumerators for filling up schedules or assisting respondents to fill up schedules and as such enumerators should be very carefully selected. The enumerators should be trained to perform their job well and the nature and scope of the investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule.

Enumerators should be intelligent and must possess the capacity of cross examination in order to find out the truth. Above all, they should be honest, sincere, and hardworking and should have patience and perseverance. This method of data collection is very useful in extensive enquiries and can lead to reliable results. It is, however, very expensive and is usually adopted in investigations conducted by governmental agencies or by some big organizations. Population census all over the world is conducted through this method.

Some other methods of data collection particularly used by big business houses in modern times are the following (Jenkins, 1975).

1. Warranty cards: Warranty cards are usually postal sized cards which are used by dealers of consumer durables to collect information regarding their products. The information sought is printed in the form of questions on the 'warranty cards' which is placed inside the package along with the product with a request to the consumer to fill in the card and post it back to the dealer.

2. Distributor or store audits: Distributor or store audits are performed by distributors as well as manufactures through their salesmen at regular intervals. Distributors get the retail stores audited through salesmen and use such information to estimate market size, market share, and seasonal purchasing pattern and so on. The data are obtained in such audits not by questioning but by observation. For instance, in case of a grocery store audit, a sample of stores is

visited periodically and data are recorded on inventories on hand either by observation or copying from store records. Store audits are invariably panel operations, for the derivation of sales estimates and compilation of sales trends by stores are their principal 'raison detre'. The

principal advantage of this method is that it offers the most efficient way of evaluating the effect on sales of variations of different techniques of in-store promotion.

3. Pantry audits: Pantry audit technique is used to estimate consumption of the basket of goods at the consumer level. In this type of audit, the investigator collects an inventory of types, quantities and prices of commodities consumed (Jenkins, 1975).

Thus, in pantry audit data are recorded from the examination of consumer's pantry. The usual objective in a pantry audit is to find out what types of consumers buy certain products and certain brands, the assumption being that the contents of the pantry accurately portray consumer's preferences. Quite often, pantry audits are supplemented by direct questioning relating to reasons and circumstances under which products were purchased in an attempt to relate these factors to purchasing habits. A pantry audit may or may not be set up as a panel operation, since a single visit is often considered enough to yield an accurate picture of consumers' preferences. An important limitation of pantry audit approach is that, at times, it may not be possible to identify consumers' preferences from the audit data alone, particularly when promotion devices produce a marked rise in sales.

4. Consumer panels: An extension of the pantry audit approach on a regular basis is known as 'consumer panel', where a set of consumers are arranged to come to an understanding to maintain detailed daily records of their consumption and the same is made available to investigator on demands. In other words, a consumer panel is essentially a sample of consumers who are interviewed repeatedly over a period. Mostly consume panels are of two types viz., the

transitory consumer panel and the continuing consumer panel. A transitory consumer panel is set up to measure the effect of a phenomenon.

Usually such a panel is conducted on a before-and-after-basis. Initial interviews are conducted before the phenomenon takes place to record the attitude of the consumer. A second set of interviews is carried out after the phenomenon has taken place to find out the consequent changes that might have occurred in the consumer's attitude. It is a favorite tool of advertising and of social research. A continuing consumer panel is often set up for an indefinite period with a view to collect data on a aspect of consumer behavior over time, generally at periodic intervals or may be meant to serve as a general-purpose panel for researchers on a variety of subjects. Such panels have been used in the area of consumer expenditure, public opinion and radio and TV listenership among others. Most of these panels operate by mail. The representativeness of the panel relative to the population and the effect of panel membership on the information obtained after the two major problems associated with the use of this method of data collection.

5. Use of mechanical devices: The use of mechanical devices has been widely made to collect information by way of indirect means (Jenkins, 1975). Eye camera, Pupillometric camera, Psychogalvanometer, Motion picture camera and Audiometer are the principal devices so far developed and commonly used by modern big business houses, mostly in the developed world for the purpose of collecting the required information. Eye cameras are designed to record the focus of eyes of a respondent on a specific portion of a sketch or diagram or written material. Such an information is useful in designing advertising material. Pupillometric cameras record

dilation of the pupil as a result of a visual stimulus. The extent of dilation shows the degree of interest aroused by the stimulus. Psychogalvanometer is used for measuring the extent of body excitement as a result of the visual

stimulus. Motion picture cameras can be used to record movement of body of a buyer while deciding to buy a consumer good from a shop or big store. Influence of packaging or the information given on the label would stimulate a buyer to perform certain physical movements which can easily be recorded by a hidden motion picture camera in the shop's four walls. Audiometers are used by some TV concerns to find out the type of programmed as well as stations preferred by people. A device is fitted in the television instrument itself to record these changes. Such data may be used to find out the market share of competing television stations.

6. Projective techniques: Projective techniques (or what are sometimes called as indirect interviewing techniques) for the collection of data have been developed by psychologists to use projections of respondents for inferring about underlying motives, urges, or intentions which are such that the respondent either resists to reveal them or is unable to figure out himself. In projective techniques the respondent in supplying information tends unconsciously to project his own attitudes or feelings on the subject under study. Projective techniques play an important role in motivational researches or in attitude surveys (Jenkins, 1975).

The use of these techniques requires intensive specialized training. In such techniques, the individual's responses to the stimulus-situation are not taken at their face value. The stimuli may arouse many kinds of reactions. The nature of the stimuli and the way in which they are

presented under these techniques do not clearly indicate the way in which the response is to be interpreted. The stimulus may be a photograph, a picture, an inkblot and so on.

Responses to these stimuli are interpreted as indicating the individual's own view, his personality structure, his needs, tensions, etc. in the context of some pre-established psychological conceptualization of what the individual's responses to the stimulus mean.

Brief of **some of the important projective techniques** are given below (Mertens, 2017).

(i) Word association tests: These tests are used to extract information regarding such words which have maximum association. In this sort of test the respondent is asked to mention the first word that comes to mind, ostensibly without thinking, as the interviewer reads out each word from a list. If the interviewer says cold, the respondent may say hot and the like ones. The general technique is to use a list of as many as 50 to 100 words. Analysis of the matching words supplied by the respondents indicates whether the given word should be used for the contemplated purpose. The same idea is exploited in marketing research to find out the quality that is mostly associated to a brand of a product. This technique is frequently used in advertising research.

(ii) Sentence completion tests: These tests happen to be an extension of the technique of word association tests. Under this, informant may be asked to complete a sentence (such as: persons who wear Khadi are...) to find association of Khadi clothes with certain personality characteristics. Several sentences of this type might be put to the informant on the same subject. This technique permits the testing not only of words (as in case of word association tests), but of

ideas as well and thus, helps in developing hypotheses and in the construction of questionnaires. This technique is also quick and easy to use, but it often leads to analytical problems, particularly when the response happens to be multidimensional.

(iii) Story completion tests: Such tests are a step further wherein the researcher may contrive stories instead of sentences and ask the informants to complete them. The respondent is given just enough of story to focus his attention on a given subject and he is asked to supply a conclusion to the story (Mertens, 2017).

(iv) Verbal projection tests: These are the tests wherein the respondent is asked to comment on or to explain what other people do. For example, why do people smoke? Answers may reveal the respondent's own motivations.

(v) Pictorial techniques: There are several pictorial techniques. The important ones are :(a) Thematic apperception test (T.A.T.), (b) Rosenzweig test, (c) Rorschach test, (d) Holtzman Inkblot Test (HIT): and (e) Tomkins-Horn picture arrangement test:

(a) Thematic apperception test (T.A.T.): The TAT consists of a set of pictures (some of the pictures deal with the ordinary day-to-day events while others may be ambiguous pictures of unusual situations) that are shown to respondents who are asked to describe what they think the pictures represent. The replies of respondents constitute the basis for the investigator to draw inferences about their personality structure, attitudes, etc.

(b) Rosenzweig test: This test uses a cartoon format wherein we have a series of cartoons with words inserted in 'balloons' above. The respondent is asked to put his own words in an empty

balloon space provided for the purpose in the picture. From what the respondents write in this fashion, the study of their attitudes can be made.

(c) Rorschach test: This test consists of ten cards having prints of inkblots. The respondents are asked to describe what they perceive in such symmetrical inkblots and the responses are interpreted based on some pre-determined psychological framework. This test is frequently used but the problem of validity still remains a major problem of this test.

(d) Holtzman Inkblot Test (HIT): This test from W.H. Holtzman is a modification of the Rorschach Test explained above. This test consists of 45 inkblot cards (and not 10 inkblots as we find in case of Rorschach Test) which are based on color, movement, shading and other factors involved in inkblot perception. Only one response per card is obtained from the subject (or the respondent) and then responses of a subject are interpreted at three levels of form appropriateness.

Form responses are interpreted for knowing the accuracy (F) or inaccuracy (F–)of respondent's percepts; shading and color for ascertaining his affectional and emotional needs; and movement responses for assessing the dynamic aspects of his life (Mertens, 2017).

(e) Tomkins-Horn picture arrangement test: This test is designed for group administration. It consists of twenty-five plates, each containing three sketches that may be arranged in different ways to portray sequence of events. The respondent is asked to arrange them in a sequence which he considers as reasonable. The responses are interpreted as providing evidence confirming certain norms, respondent's attitudes, etc.

(vi) Play techniques: Under play techniques subjects are asked to improvise or act out a situation in which they have been assigned various roles. The researcher may observe such traits as hostility, dominance, sympathy, prejudice or the absence of such traits. These techniques have been used for knowing the attitudes of younger ones through manipulation of dolls. Dolls representing different racial groups are usually given to children who can play with them freely. The way children organize dolls would indicate their attitude towards the class of persons represented by dolls. This is also known as doll-play test and is used frequently in studies pertaining to sociology. The choice of color, form, words, the sense of orderliness and other reactions may provide opportunities to infer deep-seated feelings (Mertens, 2017).

(vii) Quizzes, tests and examinations: This is also a technique of extracting information regarding specific ability of candidates indirectly. In this procedure both long and short questions are framed to test through them the memorizing and analytical ability of candidates.

(viii) Sociometry: Sociometry is a technique for describing the social relationships among individuals in a group. In an indirect way, sociometry attempts to describe attractions or repulsions between individuals by asking them to indicate whom they would choose or reject in various situations. Thus, sociometry is a new technique of studying the underlying motives of respondents.

"Under this an attempt is made to trace the flow of information amongst groups and then examine the ways in which new ideas are diffused. There are many versions of the sociogram

pattern, and the reader is suggested to consult specialized references on sociometry for the purpose.

7. Depth interviews: Depth interviews are those interviews that are designed to discover underlying motives and desires and are often used in motivational research. Such interviews are held to explore needs, desires and feelings of respondents. The difference lies in the nature of the questions asked. Indirect questions on seemingly irrelevant subjects provide information that can be related to the informant's behavior or attitude towards the subject under study.

8. Content-analysis: Content-analysis consists of analyzing the contents of documentary materials such as books, magazines, newspapers and the contents of all other verbal materials which can be either spoken or printed. Content analysis prior to 1940's was mostly quantitative analysis of documentary materials concerning certain characteristics that can be identified and counted. But since 1950's content-analysis is mostly qualitative analysis concerning the general import or message of the existing documents. "The difference is somewhat like that between a casual interview and depth interviewing

Collection of Secondary Data

Secondary data means data that are already available i.e., they refer to the data which have already been collected and analyzed by someone else (Wilcox, 1987). When the researcher utilizes secondary data, then he has to look into various sources from where he can obtain them. In this case he is certainly not confronted with the problems that are usually associated with the collection of original data. Secondary data may either be published data or unpublished data.

Usually published data are available in (Wilcox, 1987):

- various publications of the central, state are local governments;
- various publications of foreign governments or of international bodies and their subsidiary organizations;
- technical and trade journals;
- books, magazines and newspapers;
- reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.;
- reports prepared by research scholars, universities, economists, etc. in different fields; and
- Public records and statistics, historical documents, and other sources of published information.

The sources of unpublished data are many; they may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labor bureaus and other public/ private individuals and organizations.

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be un-suitable or may be inadequate in the context of the problem which the researcher wants to study. By way of caution, the researcher, before using secondary data, must see that they possess following characteristics:

1. Reliability of data: The reliability can be tested by finding out such things about the said data (Wilcox, 1987):

(a) Who collected the data?

(b) What were the sources of data?

(c) Were they collected by using proper methods?

(d) At what time were they collected?

(e) Was there any bias of the compiler?

(f) What level of accuracy was desired? Was it achieved?

2. Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not be used by the researcher. In this context, the researcher must very carefully scrutinize the definition of various terms and units of collection used at the time of collecting the data from the primary source originally. Similarly, the object, scope and nature of the original enquiry must also be studied.

3. Adequacy of data: If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher. The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry.

From all this we can say that it is very risky to use the already available data. The already available data should be used by the researcher only when he finds them reliable, suitable and

adequate. But he should not blindly discard the use of such data if they are readily available from authentic sources and are also suitable and adequate for in that case it will not be economical to spend time and energy in field surveys for collecting information. At times, there may be wealth of usable information in the already available data which must be used by an intelligent researcher but with due precaution.

Selection of Appropriate Method for Data Collection

Thus, there are various methods of data collection. As such the researcher must judiciously select the method/methods for his own study, keeping in view the following factors:

1. Nature, scope and object of enquiry
2. Availability of funds
3. Time factor
4. Precision required

Processing and Analysis of Data

The data, after collection, must be processed and analyzed in accordance with the outline laid down for the purpose at the time of developing the research plan (Huitema, 1980).

This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis. Technically speaking, processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. The term analysis refers to the computation of certain measures along with searching for patterns of

relationship that exist among data-groups. Thus, “in the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions”. But there are persons who do not like to make difference between processing and analysis. They opine that analysis of data in a general way involves several closely related operations which are performed with the purpose of summarizing the collected data and organizing these in such a manner that they answer the research question(s). We, however, shall prefer to observe the difference between the two terms as stated here in order to understand their implications more clearly.

Processing Operations

With this brief introduction concerning the concepts of processing and analysis, we can now proceed with the explanation of all the processing operations.

1. Editing

2. Coding

3. Classification: Most research studies result in a large volume of raw data which must be reduced into homogeneous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes based on common characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes.

Classification can be one of the following two types, depending upon the nature of the phenomenon involved:

(a) Classification according to attributes:

(b) Classification according to class-intervals:

4. Tabulation: When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarizing raw data and displaying the same in compact form (i.e., in the form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows.

Tabulation is essential because of the following reasons.

- It conserves space and reduces explanatory and descriptive statement to a minimum.
- It facilitates the process of comparison.
- It facilitates the summation of items and the detection of errors and omissions.
- It provides a basis for various statistical computations.

Elements/Types of Analysis

As stated earlier, by analysis we mean the computation of certain indices or measures along with searching for patterns of relationship that exist among the data groups. Analysis, particularly in case of survey or experimental data, involves estimating the values of unknown parameters of the population and testing of hypotheses for drawing inferences. Analysis may, therefore, be

categorized as descriptive analysis and inferential analysis (Inferential analysis is often known as statistical analysis). “Descriptive analysis is largely the study of distributions of one variable. This study provides us with profiles of companies, work groups, persons and other subjects on any of a multiple of characteristics such as size. Composition, efficiency, preferences, etc.”2. this sort of analysis may be in respect of one variable (described as unidimensional analysis), or in respect of two variables (described as bivariate analysis) or in respect of more than two variables (described as multivariate analysis). In this context we work out various measures that show the size and shape of a distribution(s) along with the study of measuring relationships between two or more variables. We may as well talk of correlation analysis and causal analysis. Correlation analysis studies the joint variation of two or more variables for determining the amount of correlation between two or more variables. Causal analysis is concerned with the study of how one or more variables affect changes in another variable. It is thus a study of functional relationships existing between two or more variables. This analysis can be termed as regression analysis. Causal analysis is considered relatively more important in experimental researches, whereas in most social and business researches our interest lies in understanding and controlling relationships between variables then with determining causes per se and as such we consider correlation analysis as relatively more important. In modern times, with the availability of computer facilities, there has been a rapid development of multivariate analysis which may be defined as “all statistical methods which simultaneously analyze more

than two variables on a sample of observations". Usually the following analyses are involved when we make a reference of multivariate analysis (Huitema, 1980):

(a) Multiple regression analysis: This analysis is adopted when the researcher has one dependent variable which is presumed to be a function of two or more independent variables. The objective of this analysis is to make a prediction about the dependent variable based on its covariance with all the concerned independent variables.

(b) Multiple discriminant analysis: This analysis is appropriate when the researcher has a single dependent variable that cannot be measured but can be classified into two or more groups based on some attribute. The object of this analysis happens to be to predict an entity's possibility of belonging to a group based on several predictor variables.

(c) Multivariate analysis of variance (or multi-ANOVA): This analysis is an extension of two way ANOVA, wherein the ratio of among group variance to within group variance is worked out on a set of variables.

(d) Canonical analysis: This analysis can be used in case of both measurable and non-measurable variables for the purpose of simultaneously predicting a set of dependent variables from their joint covariance with a set of independent variables.

(e) Inferential analysis: This is concerned with the various tests of significance for testing hypotheses in order to determine with what validity data can be said to indicate some conclusion or conclusions. It is also concerned with the estimation of population values. It is mainly based

on inferential analysis that the task of interpretation (i.e., the task of drawing inferences and conclusions) is performed.

Statistics in Research

The role of statistics in research is to function as a tool in designing research, analyzing its data and drawing conclusions there from. Most research studies result in a large volume of raw data which must be suitably reduced so that the same can be read easily and can be used for further analysis (Freeman, 1965). Clearly the science of statistics cannot be ignored by any research worker, even though he may not have occasion to use statistical methods in all their details and ramifications. Classification and tabulation, as stated earlier, achieve this objective to some extent, but we have to go a step further and develop certain indices or measures to summaries the collected/classified data. Only after this we can adopt the process of generalization from small groups (i.e., samples) to population. If fact, there are two major areas of statistics viz., descriptive statistics and inferential statistics. Descriptive statistics concern the development of certain indices from the raw data, whereas inferential statistics concern with the process of generalization. Inferential statistics are also known as sampling statistics and are mainly concerned with two major types of problems: (I) the estimation of population parameters, and (ii) the testing of statistical hypotheses.

The important statistical measures that are used to summaries the survey/research data are:

- measures of central tendency or statistical averages;

- measures of dispersion;
- measures of asymmetry (skewness);
- measures of relationship; and
- other measures.

Amongst the measures of central tendency, the three most important ones are the arithmetic average or mean, median and mode. Geometric mean and harmonic mean are also sometimes used. From among the measures of dispersion, variance, and its square root—the standard deviation is the most often used measures. Other measures such as mean deviation, range, etc. are also used. For comparison purpose, we use mostly the coefficient of standard deviation or the coefficient of variation. In respect of the measures of skewness and kurtosis, we mostly use the first measure of skewness based on mean and mode or on mean and median. Other measures of skewness, based on quartiles or on the methods of moments, are also used sometimes. Kurtosis is also used to measure the pawedness of the curve of the frequency distribution.

Amongst the measures of relationship, Karl Pearson's coefficient of correlation is the frequently used measure in case of statistics of variables, whereas Yule's coefficient of association is used in case of statistics of attributes. Multiple correlation coefficient, partial correlation coefficient, regression analysis, etc., are other important measures often used by a researcher. Index numbers, analysis of time series, coefficient of contingency, etc., are other measures that may as well be used by a researcher, depending upon the nature of the problem under study.

Handling Large Datasets

The best method and approach to handle large dataset that are generated from areas like astronomy, bioinformatics or finance, and so on are (Freeman, 1965);

Dimensionality Reduction

Before applying the specific mining task that must be performed on a dataset, several preprocessing steps can be done. The first goal of the preprocessing step is to assure the quality of the data by reducing the noisy and irrelevant information that it could contain. The second goal is to reduce the size of the dataset, so the computational cost of the discovery task is also reduced. There are two dimensions that can be taken in account when reducing the size of the dataset. The first one is the number of instances. This problem can be addressed by sampling techniques when a smaller subset of the data holds the same information that the whole dataset. Not in all application it is the case, and sometimes the specific goal of the mining process is to find specific groups of instances with low frequency, but of high value. This data could be discarded by the sampling process, making unfruitful the process. In other applications, the data is a stream, this circumstance makes more difficult the sampling process or carries the risk of losing important information from the data if its distribution changes over time. With dimensionality reduction techniques, the number of attributes of the dataset also can be addressed. There are several areas related to the transformation of a dataset from the original representation to a representation with a reduced set of features. The goal is to obtain a new dataset that preserves, up to a level, the original structure of the data, so its analysis will result in the same or equivalent patterns present in the original data. Broadly, there are two kinds of methods for reducing the attributes in a dataset, feature selection and feature extraction.

Clustering algorithms

The clustering task can be defined as a process that, using the intrinsic properties of a dataset X , uncovers a set of partitions that represents its inherent structure. It is, thus, an unsupervised task, that relies in the patterns that present the values of the attributes that describe the dataset. The partitions can be either nested, so a hierarchical structure is represented, or disjoint partitions with or without overlapping. There are several approaches to obtain a partition from a dataset, depending on the characteristics of the data or the kind of the desired partition. Broadly these approaches can be divided in:

• Hierarchical algorithms, that result in a nested set of partitions, representing the hierarchical structure of the data. These methods are usually based on a matrix of distances/similarities and a recursive divisive or agglomerative strategy.

• Partitional algorithms, that result in a set of disjoint or overlapped partitions. There is a wider variety of methods of this kind, depending on the model used to represent the partitions or the discovery strategy used. The more representative ones include algorithms based on prototypes or probabilistically models, based on the discovery of dense regions and based on the partition of the space of examples into a multidimensional grid.

Hierarchical clustering

Hierarchical methods use two strategies for building a tree of nested clusters that partitions a dataset, divisive and agglomerative. Divisive strategies begin with the entire dataset, and each iteration it is determined a way to divide the data into two partitions. This process is repeated

recursively until individual examples are reached. Agglomerative strategies iteratively merge the most related pair of partitions according to a similarity/distance measure until there is only one partition. Usually agglomerative strategies are computationally more efficient. These methods are based on a distance/similarity function that compares partitions and examples. The values of these measures for each pair of examples are stored in a matrix that is updated during the clustering process.

Conclusion

In this section, we covered how to collect data from different sources. We also explored the idea of data collection using questionnaire method. The merit of questionnaire method was given in depth. We explain in depth the method of selecting appropriate method for data collection. Lastly, we state the differences between the primary data and secondary data. The data quality meaning and data cleaning process will be our next discussion in week 3. We shall be looking at how to perform data quality and data cleaning in next module week 3.