A Gateway to all Post Graduate Courses

An MHRD Project under its National Mission on Education through ICT (NME-ICT)

Subject: Human Resource Management

Production of Courseware

- Content for Post Graduate Courses

Paper 05: Research Methodology

Module 09: Processing of Data

ज्ञान-विज्ञान विमुक्तये

Development Team

Principal Investigator:	Prof. Ipshita Bansal Department of Management Studies BPS Women University, Khanpur Kalan, Sonepat
Paper Coordinator:	Prof. S.P. Singh Faculty of Management Studies Gurukul Kangri University, Haridwar
Content Writer:	Prof. S.P. Singh Faculty of Management Studies Gurukul Kangri University, Haridwar
Content Reviewer:	Prof. H.L. Verma Vice-Chancellor Jagannath University, Bahadurgarh, Haryana

Items	Description of Module
Subject Name	Management
Paper Name	Research Methodology
Module Title	PROCESSING OF DATA
Module ID	Module 9
Pre-Requisites	Understanding the validation and processing of data
Objectives	To study the Validation, editing, coding, classification and tabulation of data
Keywords	Validation, editing, coding, classification, tabulation

Role	Name	Affiliation
Principal Investigator	Prof.Ipshita Bansal	Department of Management Studies, BPSMV, Khanpur Kalan, Sonipat
Co-Principal Investigator		
Paper Coordinator	Prof. S.P.Singh	Department of Management Studies, GKV, Haridwar
Content Writer (CW)	Prof. S.P.Singh	Department of Management Studies, GKV, Haridwar
Content Reviewer (CR)		
Language Editor (LE)		

QUADRANT –I

1. Module : Processing of Data

2. LEARNING OUTCOME:

After studying this module, you shall be able to

- Know the nature of processing of data
- Understand the fieldwork validation
- Comprehend the data editing
- Understand the coding of data
- Become aware of the classification and tabulation of data

3. Introduction

After the data has been collected it needs processing and analyzing in accordance with an outline chalked down while formulating the research plan. This is essential to ensure that all relevant data is attained for comparisons and analysis envisaged for the study. Thus, editing, coding, classification and tabulation of data to make them capable of analysis comprises the processing of data. The computation of particular measures and the search for relationships among group of data is known as data analysis.

Thus, associations or differences accepting or rejecting the initial or new hypotheses should be statistically tested to determine the validity with which data can be stated to indicate any conclusions. But certain scholars do not make any difference between processing and analysis. They think that a number of closely related operations carried out for summarizing and organizing that the research questions find their answers

4. Data Editing

Once the validation process has been accomplished, the next step is the editing of the raw data obtained. A process of examining the raw data for the study (particularly in surveys) to find out errors and omissions and to rectify them where possible, is known as data editing. The researcher should ensure himself that the data collected has every necessary element, agrees with other realities. The data is exact in terms of information recorded and responses tried to obtain. Further, the obtained responses are in the decided format, filled consistently, complete in all respects and arranged in proper sequence to make easier for coding and tabulation.

To ensure that data screening and cleaning which is essentially the requirement of the editing process, has been carried out, the researcher needs to carry out the process at two levels, at the field editing and central editing:

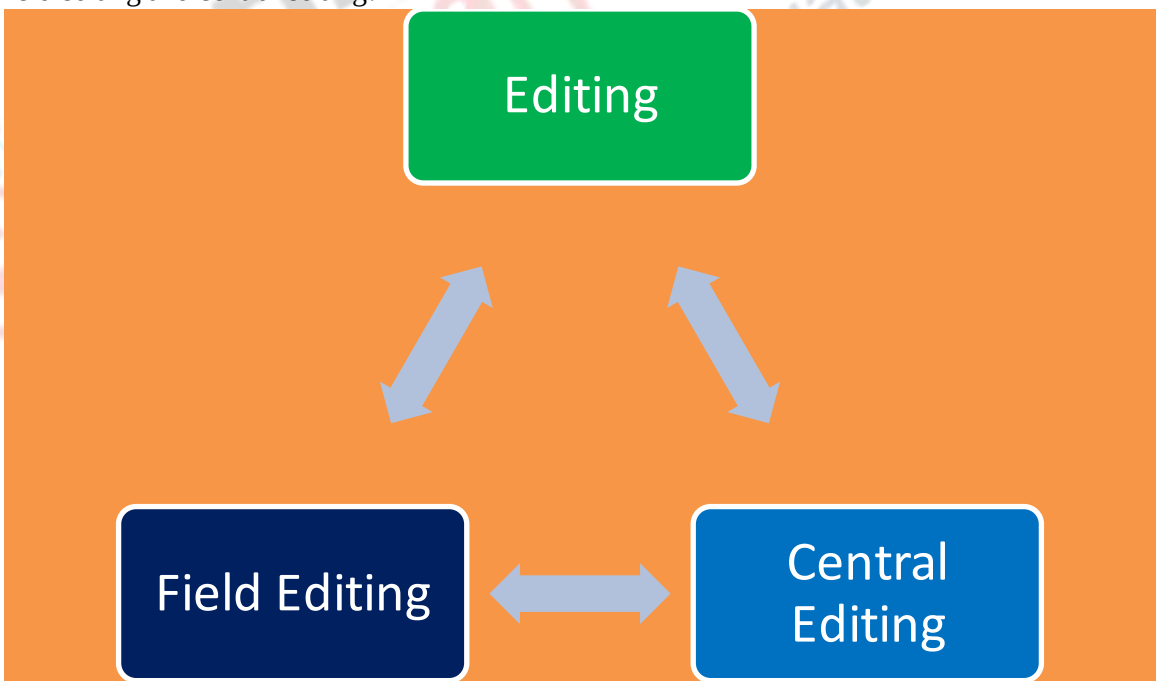

Figure 1Editing

4.1 Field-editing

In field editing the researcher examines the reporting form filled by the investigator for entirely finalizing all such forms which are written in abbreviation at the time of recording the responses. Sometimes a person's writing cannot be easily read, this kind of editing become necessary. The field editing need be carried out soon after the interview. When the

investigator finds that some questions remain unanswered while doing field editing, the investigator must resist from imagining what the respondent would have said if he had been asked the question. In quality research there is no room of self-interviewing. In large projects, the field supervisor has the responsibility for carrying out the job of field editing review and to validating the field results. Normally this may involve re-interview of some of the respondents at least on some question.

4.2 Central Editing

Central editing takes place at the time all forms lacking nothing have been received in the office. Central editing is carried out by a single editor in case of a small study and a team of editors for a large inquiry carry out a thorough editing. Editor(s) may rectify the evident errors such as an entry in the wrong space and time. Where there are inappropriate or missing replies, the proper answer is decided in the light of the other information schedule provides. Sometimes, the respondent may be contact to clarify. If the answer is inappropriate and there is no ground to determine the proper reply, the response may be deleted and in the space an entry of 'no answer' is made. From the final results, all wrong responses are deleted specifically in case of mail surveys.

Several points have to be kept in mind by the editors while performing central editing: They should well understand the instructions for the interviewers, the editing instruction and the codes furnished to the editors for the purpose of central editing. While crossing out an original entry for some reason modified or supplied. Editors should put their initials and the date of editing on each finished form.

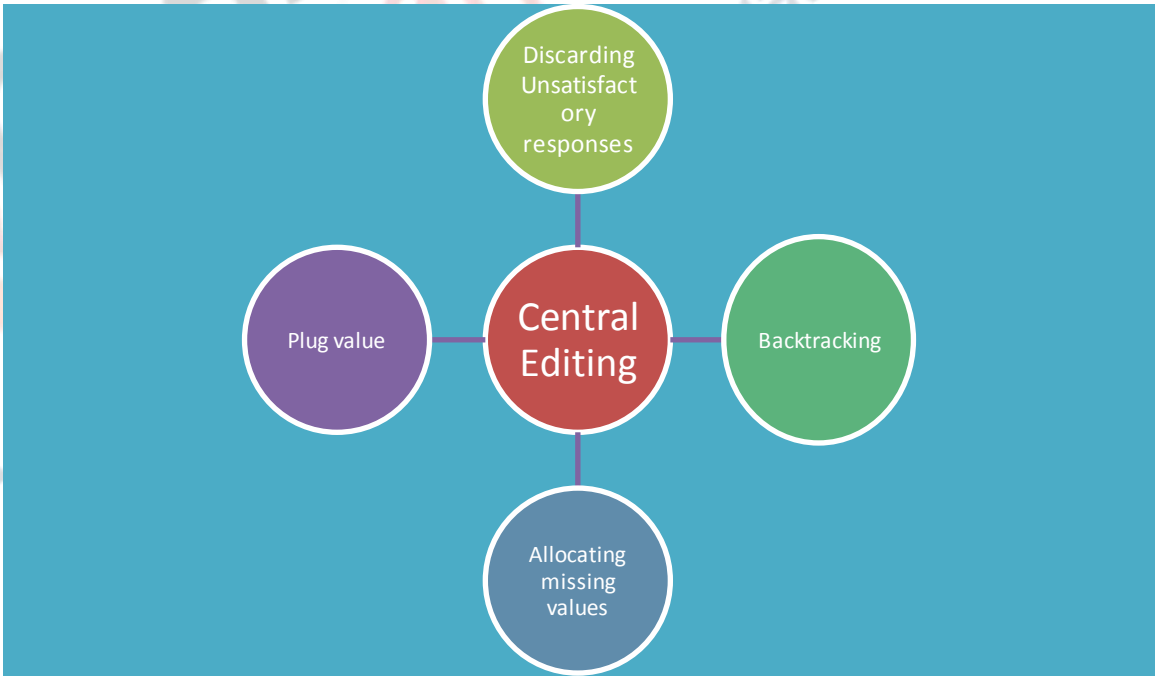

Figure 2 Essentials of Central Editing

4.2.1 Backtracking

The best and most efficient way of handling unsatisfactory responses is to return to the field, and go back to the respondents. This technique is best used for industrial surveys, where it is easier to track the respondent, who can be persuaded to give

answers to the non-response or illegible answers. In individual surveys, this becomes a little difficult as the persons locality and contact details might not be available

4.2.2 Allocating missing values

This is a contingency plan that the researcher might need to adopt in case going back to the field is not possible. Then the option might be to assign a missing value to the blanks or the unsatisfactory responses. However, this works in case:

- The number of blank or wrong answers is small
- The number of such responses per person is small
- The important parameters being studied do not have too many blanks; otherwise the sample size for those variables becomes too small for generalizations.
-

4.2.3 Plug value

When the variable under study is the key variable, then sometimes the researcher might insert a plug value. Sometimes one can plug an average or a neutral value in such cases, for example a 3 for a five-point scale. Sometimes a decision rule based upon probability could be established and the researcher might decide on a thumb rule (for example, for a yes/no question, he might decide to put 'yes' the first time he encounters missing value or no at the second and so on. Another way to handle this situation is to conduct an exploratory data analysis and see what the ratio of yes to no answers is and accordingly establish the decision rule.

Sometimes, the respondents' pattern of responses to other questions is used to extrapolate and calculate an appropriate response for the missing answer. Here, it may become a little subjective as the researcher needs to sift through the data and infer and predict the responses the person would have given had he/she answered the questions. There are statistical software and programs available to extrapolate and ascribe values for such missing responses.

4.2.4 Discarding unsatisfactory responses

If the response sheet has too many blanks/illegible or multiple responses for a single answer, the form is not worth correcting and editing. Hence, it is much better to completely discard the whole questionnaire. If too many forms are discarded then the sample for the study might become too small for an analysis or generalization, so here it is advisable to carry out another round of field visits. However, the discarding of the forms might lead to elimination from the population of the group which had a contrary or a negative opinion than the ones who completed the forms. In a research o orange juice, it happened that when the response to a product change proposition (more pulp in the drink) was studied and the completed forms were considered, they were all filled by people who liked the change, while those who did not answer all the questions had their forms rejected. Finally, when the new product was launched there were limited takers for it, as the proportion of people who did not like the drink in the studied sample was too small as compared to what existed in the actual market place.

5. Coding

Figure 3 Coding (adated from cio.com)

The process of giving numbers or symbols to responses in order to reduce the replies to limited categories appropriate to the research problem being studied is known as coding. Coding must be characterized by that there exists a category for every item of data and must have mutually exclusive meaning that a specific answer is put in only a single cell in a category set. The second guiding principle is that every category is explained in terms of only one concept. This is known as unit-dimensionality. Every efficient analysis must essentially have a stage of coding in order to reduce numerous answers to limited number of categories holding the critical information required for the analysis. The researcher must take the coding decisions when the questionnaire is designed. It make possible to pre-code the choices of the questionnaire and assist for computer tabulation. However, coding errors should be eliminated or reduced to the minimum level.

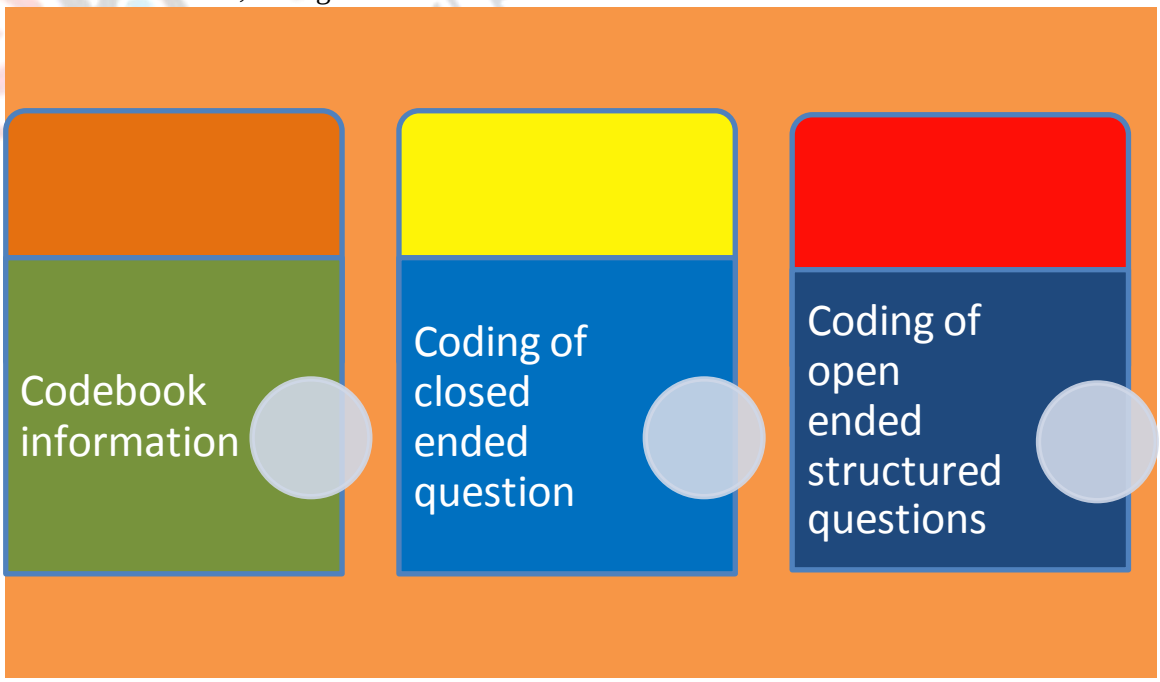

Figure 4 Coding

5.1 Codebook formulation

For simplifying and effectively managing the data entry process, it is necessary to prepare a scheme in advance for entering the data in the spreadsheet. The coding scheme for all the variables being studied is known as a code book. At the time of framing the rules categories must be decided that they are appropriate to the research objectives, are comprehensive, mutually exclusive and single variable entry. These rules become the basis of creating a code book that can be effectively used by the coders. This would generally contain information the question number, variable name, response descriptors and coding instructions and the column descriptor.

A questionnaire can have both closed-ended and open-ended questions. The process of coding the two kinds of questions is very different and requires a detailed discussion. When the questions are structure and the response categories are prescribed then one does what is called pre-coding i.e. designating numeral codes to the designed responses before administration. However, if the questions are structured and the answers are open ended and not determined in advance, one needs to decide on the codes after the administration of the survey. This is called post-coding and requires skilled interpretation and categorization of the responses into homogeneous grouped response categories and then these are assigned a numeric code.

5.2 Coding of closed-ended questions

The coding for structured questions is simple as the decision on the response categories are taken in advance. A code is assigned for every response in respect of each question and proper field and columns are specified for indicating the response codes. The responses in dichotomous question on a nominal scale can be binary.

Do you smoke? Yes= 1; no = 0

In case of ranking questions of multiple objects the person will have to make multiple columns, with column numbers equal to the number of objects to be ranked. For questions that are on a scale, the question/statement will have one column and the coding instruction would indicate numerical assignment.

5.3 Coding open-ended structured questions

The coding of open ended question is quite difficult as they are unpredictable in terms of insufficient information or a lack of hypothesis, and for this reason there are no predefined response categories. The respondents' exact answers are noted on the questionnaire. Then the researcher looks for patterns and assigns a category code. Sometimes the researcher carries out test tabulation, when he randomly looks at the answers from 20 percent of the sample data and attempts to give codes to each of the responses identified. When deciding on the codes he/she must keep the criteria of appropriateness, exhaustive categorization, mutually exclusive categories and single distribution variable as the guiding principle.

6. Classification and Tabulation

6.1 Classification:

Classification aims to reduce voluminous raw data into identical groups to derive meaningful relationships on the basis of general characteristics. Data with a general characteristic are arranged in categories in one class and thus whole data get arranged into different groups or classes. Based on the nature of the phenomenon involved, classification can be one of the two types.

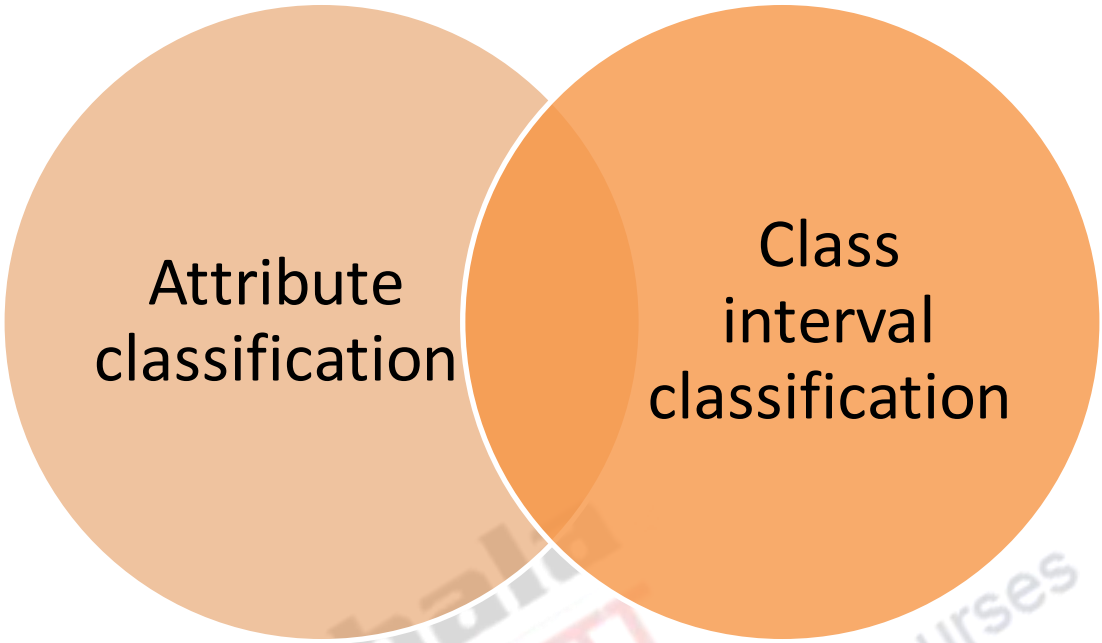

Figure 5 Classification

6.1.1 Attribute Classification

Descriptive (such as literacy, sex, honesty, etc.) and numerical (such as weight, height, income, etc.) are the two general characteristics to classify data. Descriptive characteristics mean qualitative phenomenon where it is not possible to quantify variables, the researcher simply can notice their presence or absence in a specific item. Data received on definite characteristics are known as statistics of attributes and their classification.

6.1.2 Class interval Classification

The numerical characteristics denote that quantitative phenomenon is quantitatively measured in statistical units. Data pertaining to income, production, age, weight, etc. are categorized based on class intervals. For instance, persons with incomes from Rs 1001 to Rs 4000 form one group those with income from Rs 4001 to Rs 6000 constitute another group and so on. Thus, the entire data may be categorized into different groups or 'class-intervals. Every aggregation of class-interval individually has a class limit. The dissimilarity between the two class boundaries is considered as class magnitude. There may be similar or dissimilar magnitudes in the classes. The frequency of the class is the number of items occurring in a given class. The entire classes with their respective frequencies in the form of a table are known as group frequency distribution

7. Exploratory data analysis

Once the data has been cleaned and entered in a tabular form, the researcher is advised to do a preliminary data exploration in order to assess the expected trends of the findings. Sometimes, these indicative trends may demonstrate that the data collection or instrument design is faulty and needs some corrections.

Thus, before one goes about testing the formulated hypotheses, one carries out a loosely structure exploration. Most of the exploration is done on the basis of the graphical and visual display of the

data patterns that seem to be emerging. The following are some widely used and simplistic measures of displaying data.

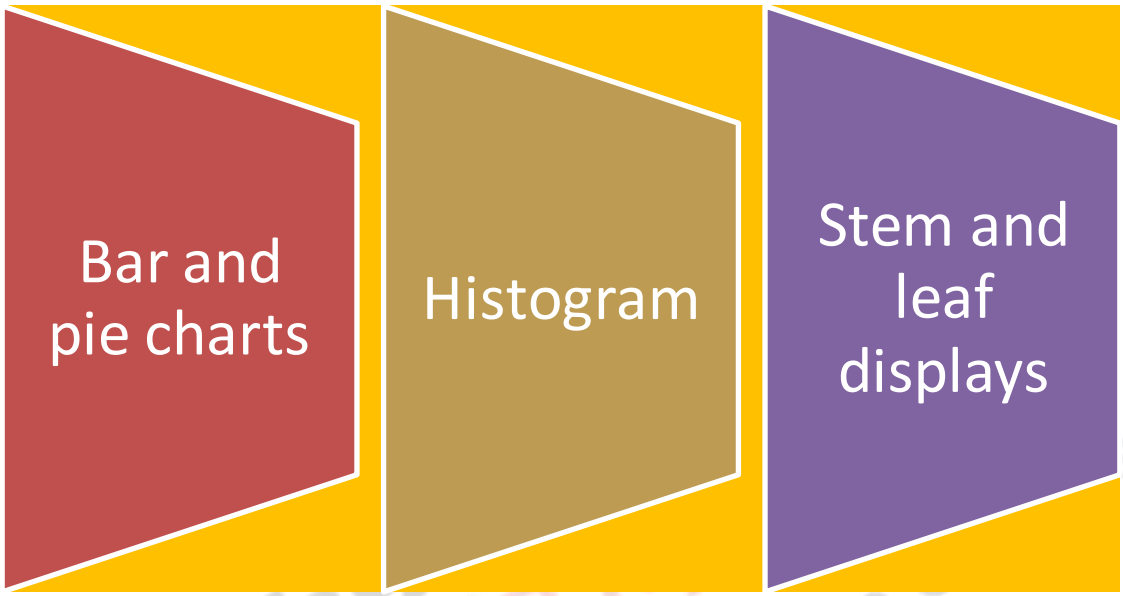

Figure 6Exploratory Data Analysis

7.1 Bar and pie charts

The data that is available as classification or demographic variable is most often on a categorical or nominal scale. Thus, the tabled data can be plotted to demonstrate the pattern of responses. For example, in a study on jeweler buying the age groups of the sample group and the occupations were as follows:

7.2 Histogram

For metric-interval and ratio scale data, the data is represented through a histogram. The representation would be able to demonstrate the distribution pattern in terms of whether it is normally distributed or demonstrates skewness.

7.3 Stem and leaf displays

This is another way of displaying the metric data. It is very easy to compute and can be done manually or with the help of Minitab. It shows individual data values in each set as against the histogram which presents only group aggregates.

It shows the pattern of responses in each interval and yet can maintain the rank order for a quick approximation of the median or quartile. Each row or line is called a stem and each value on the line is a leaf. The same data that we represented on the histogram can also be depicted on a stem and leaf display

If we look at the tabled data for the jeweler purchase in the above stem and leaf display, the decimals have been rounded off the first place and in case of two similar entries the number 13.3 has been entered twice. In fact, if one rotates the above display by 90 degrees to the left one would get the histogram. The display is showing at a glance that the sample studied was concerned with the buying of mostly 13 g items.

8. Statistical Software Packages

Researchers now find a wide array of statistical programs to assist them in both data management and data analysis. The most frequently used statistical packages are M.S.Excel, Minitab, System for System for Statistical Analysis. In M.S.Excel basic mathematical functions can be calculated. The software is easy to understand and used by most computer users. The data entered on Excel can be transported to most statistical packages for a higher level analysis. The Minitab can be used considerable easy and effectiveness in all business areas.

Summary

After the data has been collected it needs processing and analyzing in accordance with an outline chalked down while formulating the research plan. This is essential to ensure that all relevant data is attained for comparisons and analysis envisaged for the study. Thus, editing, coding, classification and tabulation of data to make them capable of analysis comprises the processing of data. A process of examining the raw data for the study (particularly in surveys) to find out errors and omissions and to rectify them where possible, is known as data editing. Central editing takes place at the time all forms lacking nothing have been received in the office. Central editing is carried out by a single editor in case of a small study and a team of editors for a large inquiry carry out a thorough editing. aaThe process of giving numbers or symbols to responses in order to reduce the replies to limited categories appropriate to the research problem being studied is known as coding. The coding consists of codebook formulation, coding of closed-end questions, of open ended structured questions. Classification aims to reduce voluminous raw data into identical groups to derive meaningful relationships on the basis of general characteristics. Classification, depending upon the nature of the phenomenon involved, can be one of the two types such as attribute classification and class interval classification. Once the data has been cleaned and entered in a tabular form, the researcher is advised to do a preliminary data exploration in order to assess the expected trends of the findings.

.