LECTURE 18 DATA EDITING, CODING, CLASSIFICATION & TABULATION

DR.GAURAV SANKALP

INTRODUCTION: DATA PROCESSING

This data preparation for research analysis is teamed as processing of data. Further selections of tools for analysis would to a large extent depend on the results of this data processing. Data processing is an intermediary stage of work between data collections and data interpretation. The data gathered in the form of questionnaires/interview schedules/field notes/data sheets is mostly in the form of a large volume of research variables. The research variables recognized is the result of the preliminary research plan, which also sets out the data processing methods beforehand. Processing of data requires advanced planning and this planning may cover such aspects as identification of variables, hypothetical relationship among the variables and the tentative research hypothesis.

The various steps in processing of data may be stated as:

- 7.2.1 Editing
- 7.2.2 Coding
- 7.2.3 Classification
- 7.2.4 Tabulation

EDITING:

Editing of data is a process of examining the collected raw data (specially in surveys) to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and/or schedules. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as completed as possible and have been well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and central editing. Field editing consists in the review of the reporting forms by the investigator for completing (translating or rewriting) what the latter has written in abbreviated and/or in illegible form at the time of recording the respondents' responses. This type of editing is necessary in view of the fact that individual writing styles often can be difficult for others to decipher. This sort of editing should be done as soon as possible after the interview,

preferably on the very day or on the next day. While doing field editing, the investigator must restrain himself and must not correct errors of omission by simply guessing what the informant would have said if the question had been asked.

Central editing should take place when all forms or schedules have been completed and returned to the office. This type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a large inquiry. Editor(s) may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule.

A t times, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final results, especially in the context of mail surveys.

Editors must keep in view several points while performing their work: They should be familiar with instructions given to the interviewers and coders as well as with the editing instructions supplied to them for the purpose. While crossing out an original entry for one reason or another, they should just draw a single line on it so that the same may remain legible. They must make entries (if any) on the form in some distinctive colour and that too in a standardised form. They should initial all answers which they change or supply. Editor's initials and the date of editing should be placed on each completed form or schedule.

CODING:

Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to be observed is that of uni-dimensionality by which is meant that every class is defined in terms of only one concept.

Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis. Coding decisions should usually be taken at the designing stage of the questionnaire. This makes it possible to pre-code the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch from the original questionnaires. But in case of hand coding some standard method may be used. One such standard method is to code in the margin with a coloured pencil. The other method can be to transcribe the data from the questionnaire to a coding sheet. Whatever method is adopted, one should see that coding errors are altogether eliminated or reduced to the minimum level.

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 on the basis of 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:

Classification according to attributes: As stated above, data are classified on the basis of common characteristics which can either be descriptive (such as literacy, sex, honesty, etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively; only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are known as statistics of attributes and their classification is said to be classification according to attributes.

Such classification can be simple classification or manifold classification. In simple classification we consider only one attribute and divide the universe into two classes— one class consisting of items possessing the given attribute and the other class consisting of items which do not possess the given attribute. But in manifold classification we consider two or more attributes simultaneously, and divide that data into a number of classes (total number of classes of final order is given by 2n, where n = number of attributes considered). Whenever data are classified according to attributes, the researcher must see that the attributes are

defined in such a manner that there is least possibility of any doubt/ambiguity concerning the said attributes.

Classification according to class-intervals: Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to income, production, age, weight, etc. come under this category. Such data are known as statistics of variables and are classified on the basis of class intervals. For instance, persons whose incomes, say, are within Rs 201 to Rs 400 can form one group; those whose incomes are within Rs 401 to Rs 600 can form another group and so on. In this way the entire data may be divided into a number of groups or classes or what are usually called, 'class-intervals.' Each group of class interval, thus, has an upper limit as well as a lower limit which are known as class limits.

The difference between the two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution.

Classification according to class intervals usually involves the following three main problems:

1. How many classes should be there?
2. What should be their magnitudes?
3. There can be no specific answer with regard to the number of classes.

The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class-intervals should be of equal magnitudes, but in some cases unequal magnitudes may result in better classification.

Hence researcher's objective judgement plays an important part in this connection. Multiples of 2, 5 and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by **H.A. Sturges**, determining the size of class interval:

$i = R/(1 + 3.3 \log N)$ where

i = size of class interval;

R = Range (i.e., difference between the values of the largest item and smallest item among the given items);

N = Number of items to be grouped.

It should also be kept in mind that in case one or two or very few items have very high or very low values, one may use what are known as open-ended intervals in the overall frequency distribution. Such intervals may be expressed like under Rs 500 or Rs 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researcher must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are to be classified.

How to choose class limits?

While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by 2) of a class-interval and the actual average of items of that class interval should remain as close to each other as possible. Consistent with this, the class limits should be located at multiples of 2, 5, 10, 20, 100 and such other figures. Class limits may generally be stated in any of the following forms:

Exclusive type class intervals: They are usually stated as follows:

10–20

20–30

30–40

40–50

The above intervals should be read as under:

10 and under 20

20 and under 30

30 and under 40

40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to the upper limit of a class are grouped in the next higher class. For example, an item whose value is exactly 30 would be put in 30–40 class intervals and not in 20–30 class intervals. In simple words, we can say that under exclusive type class intervals, the upper limit of a class interval is excluded and items with values less than the upper limit (but not less than the lower limit) are put in the given class interval.

Inclusive type class intervals: They are usually stated as follows:

11–20

21–30

31–40

41–50

In inclusive type class intervals the upper limit of a class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11–20 class intervals.

The stated upper limit of the class interval 11–20 is 20 but the real limit is 20.99999 and as such 11–20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be a discrete one (i.e., can be measured and stated only in integers), then we should adopt inclusive type classification. But when the phenomenon happens to be a continuous one capable of being measured in fractions as well, we can use exclusive type class intervals.

How to determine the frequency of each class?

This can be done either by tally sheets or by mechanical aids. Under the technique of tally sheet, the class-groups are written on a sheet of paper (commonly known as the tally sheet) and for each item a stroke (usually a small vertical line) is marked against the class group in which it falls. The general practice is that after every four small vertical lines in a class group,

the fifth line for the item falling in the same group is indicated as horizontal line through the said four lines and the resulting flower (IIII) represents five items. All this facilitates the counting of items in each one of the class groups. Alternatively, class frequencies can be determined, especially in case of large inquires and surveys, by mechanical aids i.e., with the help of machines viz., sorting machines that are available for the purpose. Some machines are hand operated, whereas other work with electricity. There are machines which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

TABULATION:

When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in some kind of 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.

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

Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulating machines or computers. In relatively large inquiries, we may use mechanical or computer tabulation if other factors are favourable and necessary facilities are available. Hand tabulation is usually preferred in case of small inquiries where the number of questionnaires is small and they are of relatively short length. Hand tabulation may be done using the direct tally, the list and tally or the card sort and count methods. When there are simple codes, it is feasible to tally directly from the questionnaire. Under this method, the codes are written on a sheet of paper, called tally sheet, and for each response a stroke is marked against the code in which it falls. Usually after every four strokes against a particular code, the fifth response is indicated by drawing a diagonal or horizontal line through the strokes. These groups of five are easy to count and the data are sorted against each code conveniently. In the listing method, the code responses may be transcribed onto a large work-sheet, allowing a line for each questionnaire. This way a large number of questionnaires can

be listed on one work sheet. Tallies are then made for each question. The card sorting method is the most flexible hand tabulation. In this method the data are recorded on special cards of convenient size and shape with a series of holes. Each hole stands for a code and when cards are stacked, a needle passes through particular hole representing a particular code. These cards are then separated and counted. In this way frequencies of various codes can be found out by the repetition of this technique. We can as well use the mechanical devices or the computer facility for tabulation purpose in case we want quick results, our budget permits their use and we have a large volume of straight forward tabulation involving a number of cross-breaks.