Data Collection

Introduction

Statistical data are the basic ‘ingredients’ of Statistics on which statistician work. A set of numbers representing records of observations is termed statistical data. The need to collect data arises in every sphere of human activity. However, that ‘Garbage in garbage out’ applies in Statistics too. Hence adequate care must be taken in the collection of data. It is a poor practice to depend on whatever data available.

Information, especially facts or numbers collected for decision making is called data. Data may be numerical or categorical. Data may also be generated through a variable.

Variable

A variable is an entity that varies from a place to place, a person to person, a trial to trial and so on. For instance, the height is a variable; domicile is a variable since they vary from person to person.

- A variable is said to be quantitative if it is measurable and can be expressed in specific units of measurement (numbers).
- A variable is said to be qualitative if it is not measurable and cannot be expressed in specific units of measurement (numbers). This variable is also called categorical variable.

The variable height is a quantitative variable since it is measurable and is expressed in a number while the variable domicile is qualitative since it is not measured and is expressed as rural or urban. It is noted that they are free from units of measurement.

Classification of Data

The data that are unorganized or have not been arranged in any way are called raw data. The ungrouped data are often voluminous, complex to handle and hardly useful to draw any vital decisions. Hence, it is essential to rearrange the elements of the raw data set in a specific pattern. Further, it is important that such data must be presented in a condensed form and must be classified according to homogeneity for the purpose of analysis and interpretation. An arrangement of raw data in an order of magnitude or in a sequence is called **array**. Specifically, an arrangement of observations in an ascending or a descending order of magnitude is said to be an **ordered array.**

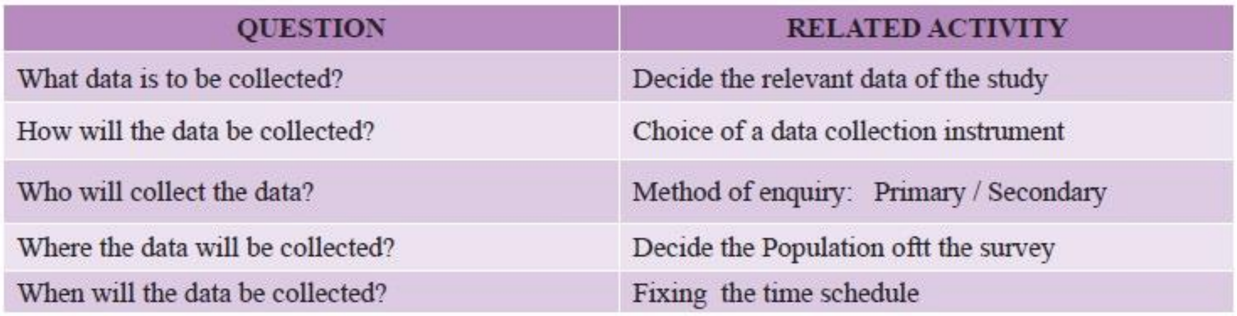
Classification is the process of arranging the primary data in a definite pattern and presenting in a systematic form. *Horace Secrist* defined classification as the process of arranging the data into sequences and groups according to their common characteristics or separating them into different but related parts. It is treated as the process of classifying the elements of observations or things into different groups or classes or sequences according to the resemblances and similarities of their character. It is also defined as the process of dividing the data into different groups or classes which are as homogeneous as possible within the groups or classes, but heterogeneous between themselves.

Objectives of Classification

- It explains the features of the data.
- It facilitates comparison with similar data.
- It strikes a note of homogeneity in the heterogeneous elements of the collected information.
- It explains the similarities which may exist in the diversity of data points.
- It is required to condense the mass data in such a manner that the similarities and dissimilarities are understood.
- It reduces the complexity of nature of data and renders the data to comprehend easily.
- It enables proper utilization of data for further statistical treatment.

Data collection process

There are five important questions to ask in the process of collecting data: What?



QUESTION	RELATED ACTIVITY
What data is to be collected?	Decide the relevant data of the study
How will the data be collected?	Choice of a data collection instrument
Who will collect the data?	Method of enquiry: Primary / Secondary
Where the data will be collected?	Decide the Population oftt the survey
When will the data be collected?	Fixing the time schedule

Data Measurement Scale

Measurement may be defined as the assignment of numbers to objects or events according to certain rules. There are generally four types of measurement scales, which are as follows.

a) Nominal scale

Nominal scale is used for measuring variables which are qualitative in nature. It is the first level of measurement where labels are assigned to the attributes of the variables in the form of number. Numbers are used as mere identifiers and do not hold any numerical value & no arithmetic operations can be drawn upon them. It only satisfies the 'Identity' property of scale of measurement. Nominal scale is the simplest scale & is also called as the 'Categorical scale' because it represents only the names or categories. It is also called as least powerful level of measurement. The only statistical analysis that can be performed on a nominal scale is frequency count. Mode is used as a measure of central tendency.

For example: -

- jersey number of players in cricket team, types of hair color, PAN number, Telephone number etc.
- Another example, what is your gender? – Male (1) or Female (2)

b) Ordinal scale

Ordinal scale is used for measuring variable which are qualitative in nature. It is the second level of measurement where labels are assigned to the variable in the form of numbers & they are arranged in a proper order. Not only the numbers but also the order of the variables is important. That's why it is called as ordinal scale. It satisfies the 'Identity' & 'Magnitude/Order' property. *Ordinal scales measure non-numeric concepts like satisfaction, happiness, discomfort, beauty etc. By giving ranks. Median or mode are used as the measures of central tendency & spearman's rank correlation.

For example,

- Order- How much happy are you with our services?

Very happy-1

Happy-2

Neutral-3

Unhappy-4

Very unhappy-5

- Another example, Ranks of students in an academic test, health status (excellent, average, poor)

Here, the order is represented but the difference between the variable is not indicated.

c) Interval scale

Interval scale is used for measuring variables which are quantitative in nature. It is the third level of measurement where labels are assigned to the variables in the form of numbers & they are arranged in a proper order with equal differences between the values. Along with the numbers & order, the difference between the values is also known. That's why it is called an Interval scale. It is an extension of ordinal scale. (i. e., it possesses the property of identity, order & equal intervals). Arithmetic operations like addition & subtraction can be performed on the variables but not multiplication & division and hence, ratios can't be calculated. Interval scales don't have a true zero meaning negative values also exist. Like -10 degree Celsius temperature. Mean, median, mode is used as the measure of central tendency. And standard deviation and range are used as the measures of dispersion. For example, a temperature scale where difference between 60 & 70 degree Celsius is same as that of the difference between 20 &30 degree Celsius

d) Ratio scale

Ratio scale is used for measuring variables which are quantitative in nature. It is the fourth level of measurement which possesses all the attributes of an interval scale along with the property of absolute zero. Arithmetic operations like addition & subtraction can be performed on the variables along with multiplication & division. Here, the ratios can be calculated. That's why it is called a ratio scale. Like, weight of ram is double of that of Shyam. Ratio scales have a true zero meaning negative values don't exist. Like there cannot be a negative weight or negative length. It is the most powerful level of measurement. Mean, median, mode, harmonic mean, geometric mean are used as the measures of central tendency. And standard deviation and coefficient of variation are used as the measures of dispersion. For example, height, weight, length, distance etc.

Summary

Feature	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Level of measurements	First	Second	Third	Fourth
Type of variable	Qualitative	Qualitative	Quantitative	Quantitative
Identity	Yes	Yes	Yes	Yes
Magnitude/order	No	Yes	Yes	Yes
Equal interval	No	No	Yes	Yes
Absolute zero	No	No	No	Yes
Central tendency	Mode	Median & mode	Mean, median & mode	Mean, median, mode, geometric & harmonic mean
Source of dispersion	…………..	………………	Standard deviation & range	Standard deviation & coefficient of variation
Arithmetic operation	…………..	……………..	Only addition & subtraction	Add, subtract, multiply & divide
Statistical tests	Non-parametric	Non-parametric	Parametric	Parametric

Types of Data

One of the major elements and basis of statistical research is data collection, where the most basic data that can be collected in this process is primary data. In other words, we can say that data is the basis of all statistical operations and primary data is the simplest of all data. Primary data is one of the 2 main types of data, with the second one being secondary data. These 2 data types have important uses in research.

Primary data

Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source—where the data originally originates from and are regarded as the best kind of data in research.

The sources of primary data are usually chosen and tailored specifically to meet the demands or requirements of research. Also, before choosing a data collection source, things like the aim of the research and target population need to be identified.

The various methods used to collect primary data are:

- Direct Method
- Indirect Method
- Questionnaire Method
- Local Correspondents Method
- Enumeration Method

1. Direct Method:

There are two methods under the direct method

(a) Personal Contact Method

As the name says, the investigator himself goes to the field, meets the respondents, and gets the required information. In this method, the investigator personally interviews the respondent either directly or through phone or through any electronic media. This method is suitable when the scope of investigation is small and greater accuracy is needed.

Merits:

- This method ensures accuracy because of personal interaction with the investigator.
- This method enables the interviewer to suitably adjust the situations with the respondent.

Limitations:

- When the field of enquiry is vast, this method is more expensive, time consuming and cumbersome.
- In this type of survey, there is chance for personal bias by the investigator in terms of asking ‘leading questions.

(b) Telephone Interviewing

In the present age of communication explosion, telephones and mobile phones are extensively used to collect data from the respondents. This saves the cost and time of collecting the data with a good amount of accuracy.

2. Indirect Method:

The indirect method is used in cases where it is delicate or difficult to get the information from the respondents due to unwillingness or indifference. The information about the respondent is collected by interviewing the third party who knows the respondent well.

Instances for this type of data collection include information on addiction, marriage proposal, economic status, witnesses in court, criminal proceedings etc. The shortcoming of this method is genuineness and accuracy of the information, as it completely depends on the third party.

3. Questionnaire Method

A questionnaire contains a sequence of questions relevant to the study arranged in a logical order. Preparing a questionnaire is a very interesting and a challenging job and requires good experience and skill.

The general guidelines for a good questionnaire:

- The wording must be clear and relevant to the study
- Ability of the respondents to answer the questions to be considered
- Avoid jargons
- Ask only the necessary questions so that the questionnaire may not be lengthy.
- Arrange the questions in a logical order.
- Questions which hurt the feelings of the respondents should be avoided.
- Calculations are to be avoided.
- It must be accompanied by the covering letter stating the purpose of the survey and guaranteeing the confidentiality of the information provided.

Editing the preliminary questionnaire

Once a preliminary draft of the questionnaire has been designed, the researcher is obligated to critically evaluate and edit, if needed. This phase may seem redundant, given all the careful thoughts that went into each question. But recall the crucial role played by the questionnaire.

Pre Test

Once the rough draft of the questionnaire is ready, pretest is to be conducted. This practice of pretest often reveals certain short comings in the questions, which can be modified in the final form of the questionnaire. Sometimes, the questionnaire is circulated among the competent investigators to make suggestions for its improvement. Once this has been done and suggestions are incorporated, the final form of the questionnaire is ready for the collection of data.

Advantages:

- In a short span of time, vast geographical area can be covered.
- It involves less labor.

Limitations:

- This method can be used only for the literate population.
- Some of the mailed questionnaires may not be returned.
- Some of the filled questionnaires may not be complete.
- The success of this method depends on the nature of the questions and the involvement of the respondents.

4. Local Correspondents Method

In this method, the investigator appoints local agents or correspondents in different places. They collect the information on behalf of the investigator in their locality and transmit the data to the investigator or headquarters. This method is adopted by newspapers and government agencies. This method is economical and provides timely information on a continuous basis. It involves high degree of personal bias of the correspondents.

5. Enumeration method:

In this method, the trained enumerators or interviewers take the schedules themselves, contact the informants, get replies, and fill them in their own handwriting. Thus, schedules are filled by the enumerator whereas questionnaires are filled by the respondents. The enumerators are paid honorarium. This method is suitable when the respondents include illiterates. The success of this method depends on the training imparted to the enumerators.

Secondary data

Secondary data is collected and processed by some other agency, but the investigator uses it for his study. They can be obtained from published sources such as government reports, documents, newspapers, books written by economists or from any other source., for example websites. Use of secondary data saves time and cost. Before using the secondary data, scrutiny must be done to assess the suitability, reliability, adequacy, and accuracy of the data.

Merits:

- It saves time and cost.
- If specially trained persons collect it, the quality of secondary data is better.
- It helps to make primary data collection more specific since with the help of secondary data, we can make out what are the gaps and deficiencies and what additional information needs to be collected.
- It helps to improve the understanding of the problem.
- It provides a basis for comparison for the data that is collected by the researcher.

Limitations:

- Accuracy of secondary data is not known.
- Data may be outdated.

Precaution in using secondary data

The following are the main precautions that should be taken before using secondary data.	
(1) Reliable agency	● We must ensure the agency that has published the data should be reliable.
(2) Suitability for the purpose of an enquiry	● The Investigator must ensure that the data is suitable for the purpose of the present enquiry. ● The suitability of the data is determined by investigating the nature, objectives, time of collection, etc. of the secondary data.
(3) Adequacy and accuracy to avoid the impact of bias	● It is necessary to use adequate data to avoid biases and prejudices leading to incorrect conclusions.
(4) Method of collecting the data used	● The investigator should also ascertain as to what method was used in collecting the data. ● Sampling methods may be biased depending upon the mode of selection of samples. ● All these should be ascertained before making use of the secondary data.

Difference between Primary data and Secondary data

Parameters of Comparison	Primary Data	Secondary Data
Definition	It is the crude form of all the data.	It is a refined form of data.
Source	It can be collected using various methods like interviews, experiments, etc.	It can be obtained from the internet, journals, etc.
Authenticity	It is very authentic in relation to the topic concerned.	It may be biased. It depends on the biases of the researcher.
Cost of collection	It is very costly to collect such data.	It costs very little or nothing.
Purpose	The primary purpose of the data is to add new knowledge.	It is a manipulated form of data and just tells the same story from a different perspective.

Census Method

The census method is also called complete enumeration method. In this method, information is collected from every individual in the statistical population. Census of Nepal is one of the best examples. It is carried out once in every ten years. An enquiry is carried out, covering each and every house in Nepal. It focuses on demographic details. They are collected and published by the Central Bureau of Statistics Nepal.

Appropriateness of this method:

The complete enumeration method is preferable provided the population is small and not scattered. Otherwise, it will have the following disadvantages.

Disadvantages:

- It is more time consuming, expensive and requires more skilled and trained investigators.
- More errors creep in due to the volume of work.
- Complete enumeration cannot be used if the units in the population destructive in nature. For example, blood testing, testing whether the rice is cooked or not in a kitchen,
- When area of the survey is very large and there is less knowledge about the population, this method is not practicable. For example, the tiger population in Nepal, number trees in a forest cannot be enumerated using census method.

Sampling method:

In view of all these difficulties one has to resort to sampling methods for collecting the data.

Sample is small proportion of the population taken from the population to study the characteristics of the population. By observing the sample one can make inferences about the population from which it is taken.

Sampling is a technique adopted to select a sample. The sample must represent or exactly duplicate the characteristics of the population under study. In such case that sample is called as a representative sample. The sampling method used for selecting a sample is important in determining how closely the sample resembles the population, in determining.

Sampling unit is the basic unit to be sampled from the population which cannot be further subdivided for the purpose of sampling. Head of the house is the sampling unit for the household survey. In the study to know the average age of a class, student is the sampling unit.

Sampling frame to adopt a sampling procedure it is precisely about sampling necessary to prepare a list such that there exists, one to one that “one grain suffices correspondence between sampling units and numbers. Such a list or map is called sampling frame. A list of villages in a district, Student list of +1 and +2 students in the above said example, A list of houses in a household survey etc.

Sample size is the number of units in the sample.

Merits and Limitations of Sampling

The prime objective of the sampling is to get the representative sample which will provides the desired information about the population with maximum accuracy at a given cost.

Merits

- Cost: Expenditure on conducting the survey is less compared to complete enumeration.
- Time: The consumption of time is relatively less in a sample study than potentially generated voluminous data.
- Accuracy: It is practically proved that the results based on representative samples more reliable than the complete enumeration.
- In the case of destructive type situations, sampling method is the only way.

Limitations

- Accuracy depends on the honesty of the investigator
- There is possibility for sampling error.