

## Chapter 9

### Surveys

Surveys are useful in investigating the knowledge attitude, behaviour or current health problems in a population. Such data might be used to propose theories or hypotheses about the nature and causes of illness, as well as utilization rates for services. In most countries government agencies and institutions regularly conduct surveys on samples of population for building up an information base about the general economic, social and health situation in the country, or to monitor patterns of consumption.

Health surveys are commonly used for the following:

- 1) To establish the attitudes, opinions or beliefs of people concerning health related issues. The data collection techniques often include questionnaires and interviews.
- 2) To study characteristics of populations on health related variables, like utilization of health care, prevalence of illness (e.g. goitre, xerophthalmia, and so on), or of health problems e.g. prevalence of hypertension, diabetes, emotional problems, types of disability, or drug use patterns.
- 3) To collect information about the demographic characteristics (age, sex, income etc) of populations.

Surveys like the ones mentioned above provide an overview of the state of health, illness, and treatment patterns in a given community. One can thereby gain insights into issues such as the prevalent causes of death, or health needs of the population. Significant patterns in the data may be observed indicating important differences or relationships.

#### **Descriptive Surveys**

These mainly describe the phenomenon of interest and its observed associations in order to estimate certain population parameters (e.g. the prevalence of diabetes in a community), for testing hypotheses (e.g. diabetes is more common in certain ethnic groups), and for generating hypotheses about possible cause and effect. They can range from analysis of routine statistics to a proper cross-sectional survey which describes a phenomenon and examines associations between the variables of interest. The hypotheses generated by such studies can then be further tested by analytical and experimental studies later.

#### **Analytical Surveys**

Surveys carried out longitudinally collect data at several points in time and help to suggest the direction of cause and effect association. Trend and Panel designs are all variations of longitudinal studies and are described under cross-sectional studies in Chapter 5.

## Comparison of two or more naturally occurring groups

A variety of comparisons can be made using survey data. However, for making valid comparisons a number of variables may need to be controlled, like for example, age, sex, educational background etc. All that the researcher does is to measure differences between naturally occurring groups. For example, a comparison of the prevalence rates of diarrhoea between two village communities, or of infant mortality rates between different geographical locations in a health district. If significant differences are found a variety of hypotheses may be proposed for further studies. However, one should bear in mind that in comparison studies groups may differ with regard to many variables other than those chosen by the researcher.

It is often not possible to determine causation with certainty in natural comparison studies in contrast to experimental studies. However, investigators can still gain crucial information about the differences between groups on clinically relevant matters.

## Correlation Studies

The aim of correlation studies is to identify interrelationships among clinically significant variables e.g. being over weight and coronary heart disease. It needs to be stressed that in such observations no evidence is being presented that one variable is **causally** related to another. There may be other variables like smoking, raised blood cholesterol levels, hypertension, type of personality, and so on, which may correlate equally with the probability of heart disease. Being overweight does not totally account for coronary artery disease. In fact, some overweight individuals do not suffer from it.

## Quasi experimental designs

If interventions are to be undertaken to reduce exposure to risk factors of a disease then reasonable evidence should be mustered to show that these factors are causally related. Quasi experimental designs is one way of obtaining such evidence. Quasi experimental designs are so called because they resemble experiments with the important difference that there is no random assignments into treatment groups. The investigator controls the time at which an intervention is introduced or withdrawn. Examples of quasi experimental designs are the time series and multiple group time series designs.

### Time-series design.

This involves repeated observations before and after a given intervention. For example, change in the prevalence rates of measles following an immunization campaign, or deaths from diarrhoea since the promotion of oral rehydration therapy. In this way changes observed following the introduction of an intervention may represent the effects of the intervention.

In an investigation of change in the prevalence of malaria after the introduction of pyrethrin impregnated bed nets one may carry out the following procedures:

- 1). Select an appropriate community for study.
- 2). Define clearly the dependent variable "prevalence of malaria".
- 3). Introduce the intervention.
- 4). Monitor the outcome variable over a period of time. It is essential to make observations both before and after the introduction of impregnated bed nets.

There is, however, a problem of spurious results due to self induced changes in life style by the subjects like use of chemoprophylaxis or clearing of mosquito breeding sites as a result of the interest generated by the campaign.

### **Multiple Group Time Series Design.**

Here comparison is made between two or more groups of subjects. For example, in the example of the malaria study quoted above, the investigator may select a community which is closely matched to the index community under study. Then, for

Community A (Control) has no health promotion activity.

Community B. (Index). Introduce the health promotion programme.

Even then problems of validity remain. There is no guarantee that communities A and B were equivalent in all relevant factors, or that no change occurred in them during the study.

Because research workers have less control over conditions in the case of naturally occurring groups, correlational and time-series designs do not always give unequivocal explanations about the cause. Hence researchers use evidence from a variety of investigations using different types of designs to evaluate their hypotheses.

## **Estimating the burden of disease**

In many developing countries limited information is often available for making quantitative assessments of the extent of different disease problems either nationally or districtwise. Such information is necessary not only to assist in the allocation of health resources but also to determine research priorities. Health statistics in one form or another are collected by clinics and hospitals, and these then get collated centrally. Unfortunately, the reliability of the basic data is often questionable and is therefore put to little use. A vicious cycle is created where the collectors of the data do not see the value of its collection and hence have little incentive to improve its reliability. The situation is not

totally bleak. There are some countries where such data are reasonably well collected and utilized, particularly for the surveillance of epidemic diseases.

There is a continuing need for detailed surveys of samples of the population in a defined community or geographic area to determine its health problems, either with respect to a particular disease or, to the range of disease problems it experiences. Such surveys may also be required to determine the patterns of social habits and practices that affect health status. The simplest epidemiological survey that may be directed towards these ends is the cross-sectional study. Random sampling can often pose a formidable logistical problem, and an alternative approach is cluster sampling. (For details of cluster sampling see Chapter 3).

A simple method of determining disease incidence is to conduct repeated cross-sectional studies at fixed lengths of time (Trend Design - as mentioned above). The interval between the surveys will depend on the factors under study. If a disease is being studied which has repeated short episodes (for example, diarrhoea), the interval would be made short to provide for accurate recall. On the other hand, if child growth is being studied, a longer interval would do.

To measure disease incidence rates or to study the natural history of a disease through repeated cross-sectional surveys it would be necessary to link individuals between the surveys. Such studies may be difficult to conduct and interpret in areas where there is high migration or seasonal movement of some members of the population. Also it is likely that the disease experience of those who migrate will be different from those who do not. It is also difficult to distinguish satisfactorily whether those who are absent at the time of the resurvey are just away or have died. The cost of studies in which it is planned to link individuals between surveys is likely to be considerably greater than of those without such linkages. Secondly, the analysis of data on linked records is much more complicated than is the case with simple non-linked surveys.

A practical problem in the conduct of some studies is concerning the sampling unit. In a study about immunization coverage it is usual to randomize individuals and thus the sampling unit is the individual. On the other hand in a study of the effectiveness of malaria control in communities by either promoting the use of chemo prophylaxis or of pyrethrin impregnated bed nets the sampling unit is the community since the intervention is at the community level.

## **Use of Secondary Data**

Often an existing data set may be available for answering a given research question. The data may have been collected as part of routine surveys or in connection with some other unrelated research. The advantages are of speed and economy, especially when an investigation is being planned but financial support for the research has not been secured. The main limitation is that the decisions about data collection, the variables to be measured, methods of measurement and recording, as well as about quality control are not made by the researcher. In all likelihood the variables measured may turn out to be not exactly the same as those which the researcher would have chosen, or the scales may be different.

There are two main types of secondary data viz. - individual and aggregate. Sources of **individual data** comprise previous research, medical records of patients, personnel files, death certificates, actuarial data, and so on. For a given data set associations between characteristics can be measured among individual members of the study population. In the case of **aggregate data** information is available only for groups. Sources of aggregate data are the worldwide data published by various UN agencies, national census data, hospital discharge data and so on.. Here again associations between variables can be measured, but only among groups. A major pitfall to be aware of is that associations observed in the aggregate do not necessarily hold true for the individual. This is referred to as the **Ecological Fallacy**. The other difficulty with aggregate data is the susceptibility to confounding. Groups of people, separated by time or distance tend to differ from one another in many ways which are not necessarily all causally related.

The use of secondary data for research can be considered in two ways viz.

- (i). Begin with a research hypothesis and search for an appropriate data set
- or
- (ii). Begin with a data set and look for new research questions that might be answered.

### **Beginning with a research hypothesis**

The investigator can begin with a research question and look for a data set that can provide the answer. The investigator would need to first carry out a thorough study of the literature pertaining to the question. Then having identified a promising data set he would need to draw up a list of predictor and outcome variables whose relationship is to be investigated. It is most unlikely that the data set will contain the exact variables, or that they are measured exactly as the investigator would have liked. Compromises have to be made, and proxy variables have to be settled for e.g. residential or property tax as proxy for income status, religion as proxy for alcohol consumption, and so on.

### **Beginning with a data set**

In the alternative second approach the investigator begins with a data set and looks for questions that might be answered with it. For example, looking for new findings that were not recognized previously. In this situation the investigator identifies a data set that appears promising and familiarizes himself with the information that has been gathered. A list is prepared of all the variables that have been measured. The groups of variables whose relationship is of interest are then identified and analyzed.

In real life the two approaches merge. The experienced researcher knows what are the growth areas in the subject. When a new data set becomes available, he looks for questions that might be answered by it. Similarly, when he thinks of a new hypothesis he looks for a data set with which to test it.

## **Tertiary Analysis**

At times it is necessary to pool together data from previously reported studies pertaining to the same research topic. Pooling of data sets is especially useful when individual studies give conflicting results. Pooling helps to increase the sample size, and the results are then more reliable.

Appendix 9.1.  
**Steps in Planning a Health Survey**

1). Writing a detailed statement of the objectives of the survey.

Each objective is examined to ensure that it is achievable with the available resources.  
Information on some of the objectives may be already available.

2). Determination of the information needed to achieve the objectives. This requires a clear definition of terms, the variables to be measured, categorical classification, and methods of data collection.

3). Definition of the population on which information is to be obtained.

4). Decision whether the population would be studied as a whole or whether sampling would be done.

5). If sampling is to be done, calculation of the sample size.

6). Sampling method to be used.

7). Design, testing and validation of the questionnaire.

8). Selection and training of interviewers.

9). Collection of data.

10). Data analysis.