Chapter 10

Cross-sectional surveys

Cross-sectional surveys are studies aimed at determining the frequency (or level) of a particular attribute, such as a specific exposure, disease or any other health-related event, in a defined population at a *particular point in time*. For instance, we can carry out a cross-sectional survey to estimate the prevalence of hepatitis B infection, the prevalence of smoking or the proportion of women of childbearing age who are breast-feeding in a given population at the time of the survey (Example 10.1).

Example 10.1. *The World Fertility Surveys (WFS) were national surveys of human reproductive behaviour conducted in about 40 developing and 20 developed countries in the late 1970s. Among other aspects of reproductive behaviour, these surveys collected information on breast-feeding practices (United Nations, 1987). Table 10.1 shows the percentages of women aged 15–49 years who were breast-feeding around the time of these surveys in selected countries.*

Table 10.1. Proportion of women aged 15–49 years in selected countries who were breast-feeding at the time the World Fertility Surveys were conducted, 1975–80.[a]

Region and country	Year of survey	Sample size	Percentage of women aged 15–49 years currently breast-feeding
Africa			
Egypt	1980	8788	34.3
Ghana	1979–80	6125	37.7
Kenya	1977–78	8100	43.2
Latin America and the Caribbean			
Colombia	1976	5378	17.1
Mexico	1976	7310	19.8
Venezuela	1977	4361	15.3
Asia and Oceania			
Bangladesh	1975–76	6513	51.1
Indonesia	1976	9155	15.9
Pakistan	1975	4996	40.5

[a] Data from United Nations (1987)

In this type of study, subjects are contacted at a fixed point in time and relevant information is obtained from them. On the basis of this information, they are then classified as having or not having the attribute of interest.

Example 10.2. *In the World Fertility Surveys, breast-feeding practices were examined in relation to socioeconomic factors such as mother's education (Table 10.2).*

Table 10.2.
Mean duration of breast-feeding (months) by mother's years of schooling in selected countries. World Fertility Surveys, 1975–80.[a]

Region and	**Year of**	**Years of schooling**			
Country (sample size)	**survey**	**Zero**	**1–3**	**4–6**	**7+**
Africa					
Egypt (8788)	1980	21.2	19.5	16.3	10.2
Ghana (6125)	1979–80	21.3	n.a.	19.2	15.7
Kenya (8100)	1977–78	19.6	17.4	15.2	12.5
Latin America and the Caribbean					
Colombia (5378)	1976	11.9	11.4	8.3	5.3
Mexico (7310)	1976	12.9	10.9	8.3	3.8
Venezuela (4361)	1977	11.6	10.0	6.7	3.5
Asia and Oceania					
Bangladesh (6513)	1975–76	34.4	30.4	n.a.	n.a.
Indonesia (9155)	1976	28.4	27.0	24.7	13.7
Pakistan (4996)	1975	22.0	n.a.	19.8	n.a.

[a] Data from United Nations (1987)

n.a. = data not available because of small sample sizes.

In some instances, cross-sectional surveys attempt to go further than just providing information on the frequency (or level) of the attribute of interest in the study population by collecting information on both the attribute of interest and potential risk factors. For instance, in a cross-sectional survey conducted to estimate the prevalence of hepatitis B in a given population, it is also possible to collect data on potential risk factors for this condition such as socioeconomic status, intravenous drug use, sexual behaviour, etc.

Example 10.3. *A national survey was conducted in the USA in 1966 to assess the prevalence of smoking, and attitudes and beliefs towards the use of tobacco and other related variables. The questionnaire included, among others, questions on the following topics: smoking behaviour (past and present); attempts to stop and/or cut down cigarette smoking; self-estimation of future smoking behaviour; beliefs about ability to change, and willingness to change; rationale for cigarette smoking behaviour; attitudes and beliefs about smoking as a health hazard in general, and to respondents in particular; gratification derived from smoking; and social pressures for continuation or cessation (US Department of Health, Education, and Welfare, 1969).*

In Example 10.2, breast-feeding duration was examined by years of schooling of the mother. In all countries where the comparison could be made, breast-feeding duration decreased consistently with increasing educational level of the mother.

Cross-sectional surveys are also useful in assessing practices, attitudes, knowledge and beliefs of a population in relation to a particular health-related event (Example 10.3). The results from these surveys not only give an indication of the magnitude of the problem in a particular population at a particular point in time, but also provide a basis for designing appropriate public health measures (e.g., health education campaigns).

Surveys are also a valuable method of obtaining information on the patterns of morbidity of a population, as well as on the utilization of preventive and curative health services (Example 10.4). Their results help health planners to establish health priorities.

Example 10.4. *The Danfa Comprehensive Rural Health and Family Planning Project was set up to assess health care and family-planning delivery systems in southern Ghana. As part of this project, a baseline household morbidity cross-sectional survey was undertaken in the study area to provide data on patterns of illness and disability, amount of work lost and use of health care services during the two-week period preceding the survey (Belcher* et al.*, 1976).*

10.1 Target population and study sample

As for the other types of epidemiological design, the aims of the study must be clearly established before its start. This process requires a precise definition of the attribute of interest (whether disease, exposure or any other health-related event) and of potential risk factors, and a clear consideration of the *target population*, i.e., the population to which the main results of the study will be extrapolated. For instance, if we were planning a study of the dietary habits of Seventh Day Adventists (a religious group who do not eat meat or drink alcohol), it would be necessary to decide whether to include children, recent converts to the church or those who had recently left.

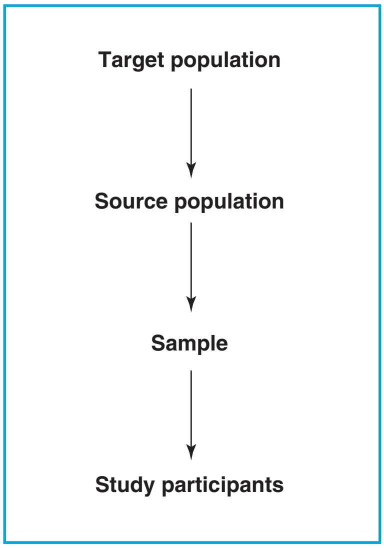

Figure 10.1.
Diagram illustrating the relationship between the target population and the study participants.

Next, a suitable *source population* needs to be identified (Figure 10.1). For practical and logistic reasons, the source population is generally more limited than the target population. For instance, although our target population comprises all Seventh Day Adventists, it would be impossible to include all of them in the study. The choice of the source population should be determined by the definition of the target population and by logistic constraints. For logistic reasons, we might decide to conduct the study in California (USA), where a large number of Seventh Day Adventists live. If this source population is small enough to be studied using the human and financial resources available, the entire population can be included. If the source population is still too large, a *representative sample* has to be selected.

10.1.1 How to select a sample?

In order to select a sample from the source population, we need to decide on the *sample design*, i.e., on the method to be used for selecting the sampling units from the population. Samples are sometimes chosen by *judgement* (i.e., what the investigator judges to be a 'balanced' or 'representative' sample) or by *convenience* (i.e., the most easily obtained subjects such as volunteers or people who attend a clinic).

None of these methods provides any guarantee against the possibility that (conscious or unconscious) selection bias may be introduced. Some people may be more likely than others to get into the sample, and the sample will become unrepresentative. For example, clinic attenders may be different from non-attenders (as in Example 10.5).

Example 10.5. *A survey was conducted in Seoul, Republic of Korea, to determine the prevalence of* Helicobacter pylori *infection in the population of the city. The sample consisted of asymptomatic healthy adults and children who visited a health screening centre at Guro Hospital for routine health examination. The majority of the individuals served by the centre were from the middle class, with fewer private patients and families from lower socioeconomic classes (Malaty* et al.*, 1996).*

The best approach is to use *random sampling*. In this method, chance alone determines who will be included in the sample, removing any possibility of selection bias.

In order to draw a random sample from the source population, we need to have a *sampling frame*, i.e., a complete enumeration of the sampling units in the study population. The sampling unit may be an individual person, a household, or a school. Electoral registers may be a suitable sampling frame for adults but not for children. If the sampling frame is based on official statistics, some groups may be under-represented, such as recent immigrants, the homeless, and slum dwellers. A sampling frame may not exist for other groups such as gypsies and other nomad groups. In certain countries it may be impossible to enumerate everyone in the study population. As we shall see later in this chapter (Section 10.1.4), special techniques can be used in these circumstances to ensure the attainment of a representative sample.

10.1.2 Simple random sampling

The most elementary kind of random sample is a simple random sample, in which each sampling unit has an equal chance of being selected directly out of the source population.

The first step is to define who are the sampling units, i.e., the people or items (e.g., households) who are to be sampled. These units need to be defined clearly in terms of their particular characteristics. The next step is to draw up a sampling frame, i.e., a list of all the sampling units in the source population. The sampling frame should be comprehensive, complete and up-to-date, to keep selection bias to a minimum. Common examples of sampling frames include census lists and electoral registers. Once a suitable sampling frame has been identified, its sampling units should be given a number. If the source population is formed by 2000 individuals, each one should be assigned a unique number between 1 and 2000. Random number tables can then be used to select a random sample out of the total sampling units who make up the source population. First, a random starting place in the table and a random direction should be selected. Then all the sequential digits found on the table should be recorded, stopping only when the required sample size is reached (see Section 7.9.2 for an illustration of how to use tables of random numbers). Alternatively, sequences of random numbers can be generated by a calculator or a computer package. The sample will be formed by the sampling units which correspond to these random numbers.

The main feature of simple random sampling is that it is relatively simple as compared with other methods (such as those described in Sections 10.1.4 and 10.1.5). Its main limitation is that it is only practicable when the population is relatively small and concentrated in a small geographical area and where the sampling frame is complete.

Example 10.6. *A cross-sectional survey was performed on random samples of women in a high-risk area for cancer of the cervix uteri (Nuuk, Greenland) and in a low-risk area (Nykøbing Falster) of Denmark to assess the prevalence of infections by specific types of human papillomavirus (HPV) and herpes simplex virus (HSV) infection. The Danish Central Population Registry is a computerized record of everyone who was alive in 1968 or who was born in or immigrated into Denmark thereafter and includes information on vital status and emigration. A sample of 800 women aged 20–39 years, born in Greenland and residing in the municipality of Nuuk/Godthåb, was drawn at random from this population registry. Similarly, a random sample of 800 women aged 20–39 years, born in Denmark and resident in Nykøbing Falster municipality, was also drawn from the same registry (Kjaer* et al.*, 1988).*

In Example 10.6, a sample of women was selected by simple random sampling in each of the two municipalities. This method was adequate and convenient since there was a proper sampling frame (i.e., the computerized list) for each area and both populations were concentrated in relatively small geographical areas.

Although it is not usually feasible to use simple random sampling for selecting the whole sample in a survey covering a large geographical area, it is often used for the final selection of the study units (e.g., selecting households in communities, after communities have been selected) within more complex schemes as described in Section 10.1.4.

10.1.3 Systematic sampling

Sometimes it may be more convenient to draw a systematic sample rather than a simple random sample. To do this, the units must be arranged in some kind of sequence as in a directory or in a series of index cards, or houses along a street, or patients as they arrive at a clinic. Then we need to decide what fraction of the population is to be studied. Suppose, for example, we wish to select a sample of 40 from a population of size 200. This will be a 1 in 5 sample. We choose a number at random between 1 and 5—let us suppose it was 2. Then starting at the beginning of the list, we select the sampling unit number 2 and then every fifth subsequent unit. The sample will include units 2, 7, 12, 17,... and so on.

The main advantage of systematic sampling is convenience. It generally provides a good approximation to simple random sampling provided that the intervals do not correspond with any recurring pattern in the source population. If particular characteristics arise in the sampling frame at regular intervals, bias will be introduced. Consider what would happen if the population were made up of a series of married couples with the husband always listed first. Picking every fourth person would result in a sample constituted exclusively by men if one started with the first or third subject or exclusively by women if one started with the second or fourth. Similarly, every 20th house on a list of addresses or houses might be a corner house with different characteristics to the other houses.

10.1.4 Multi-stage sampling

In many situations, it is not feasible or practical to draw a simple random sample or a systematic sample from the whole source population. This is either because a sampling frame is unavailable and the effort involved in drawing one up would be too great, or because the population is dispersed over a very large geographical area. For example, it would be unrealistic to try to draw a simple random sample of 200 people from the population of an entire country. Even if a proper sampling frame did exist, most of the sample would live in different communities far away from each other, and the time and expense involved in contacting them would be prohibitive. One solution is to use *two-stage sampling* as follows:

1. The population is first divided into *clusters*, for example regions, villages or districts, and a list of these *first-stage units* (or primary sample units) drawn.

2. A random sample of first-stage units is then selected from this list.

3. In each of the selected first-stage units, a sampling frame of the *second-stage units* (e.g., households or individuals) is drawn up, and a random sample of these selected.

Example 10.7. *A large cross-sectional survey was carried out in China in 1983 at the end of the harvest season to provide information on diet and lifestyle. A multi-stage sampling procedure was used to select the participants.*

(1) Sixty-five rural counties (almost all with populations over 100 000 in 1973–75) were selected from a total of 2392 counties. The county selection was not random, but designed to produce geographical areas with a wide range of cancer rates for seven of the most common cancers and wide geographical scatter.

(2) Two communes in each of the 65 counties were selected, i.e., a total of 130 communes. Although the selection of the communes was random, a decision was made to keep all communes within four hours' travel time from the survey station (in the commune) to the county laboratory, resulting in replacement of six communes.

(3) One production brigade in each commune was randomly selected, i.e., a total of 130 production brigades.

(4) Two production teams in each of the 130 production brigades were randomly selected, i.e., a total of 260 production teams.

(5) Within each of the 260 production teams, 25 households were randomly selected from an official registry of residences (yielding a total of 100 households per county and 50 households per commune).

(6) For each household, either one male or one female aged 35–64 years of age was then invited to donate blood and to complete a questionnaire about their dietary, drinking, smoking, and reproductive histories. A total of 6500 subjects participated in the study (Chen et al., *1990).*

This strategy can be extended to several stages (*multi-stage sampling*), as in Example 10.7.

If no frame of households exists and it is not practical to create one, some selection method has to be used which ensures that the sample is as representative as possible. This usually involves two phases: a method of selecting one household to be the starting point and a procedure for selecting succeeding households after that. One possibility is to choose some central point in a town, such as the market or the central square; choose a random direction from that point (e.g., by throwing a pencil in the air and seeing which way it lands); count the number of households between the central point and the edge of town in that direction; select one of these houses at random to be the starting point of the survey. The remaining households in the sample should be selected to give a wide-spread coverage of the town. The precise method used is not too important, as long as it does not result in all the chosen households being very close to one another, and as long as the rule for selecting households after the first is simple and unambiguous, to remove the possibility of interviewers introducing bias by avoiding certain areas.

If the objective of the study is to obtain an overall estimate of the prevalence or level of an attribute across the whole target population, it is sensible to select the various geographical areas in such a way that the probability of selection is proportional to the size of their population. For instance, a town with a population of 200 000 should stand ten times the chance of being selected as a town with a population of 20 000. A similar number of individuals or households should then be taken from both small and large towns. This *probability proportional to size sampling* approach ensures that individuals (or households) in both large and small towns stand an equal probability of being selected at the start of the sampling procedure.

The advantages of multiple-stage sampling are obvious in terms of costs and time. Thus, should all samples be obtained by selection of convenient clusters? One drawback of this method is that in many situations the clusters are likely to be formed by sets of individuals that are more homogeneous than the population as a whole. For instance, people living in the same neighbourhood or village are likely to be similar in terms of their lifestyle characteristics. If this is the case, individuals in a sample of neighbourhoods provide less information than a sample of similar size obtained from the whole study population.

10.1.5 Stratified random sampling

A stratified random sample involves dividing the population into distinct subgroups according to some important characteristics, such as sex, age or socioeconomic status, and selecting a random sample out of each subgroup. Each subgroup is known as a *stratum* (plural*: strata*), and a separate random sample (simple or multi-stage) is selected in each one.

Example 10.8. *The seroprevalences of immunoglobulin G (IgG), M (IgM), and A (IgA) antibodies to* Helicobacter pylori *were assessed by enzyme-linked immunosorbent assay techniques in a survey conducted in the western part of Copenhagen County (Denmark). In 1982, an age- and sex-stratified sample consisting of 4807 men and women born in the years 1922, 1932, 1942, and 1952 (i.e., aged 30, 40, 50 or 60 years) and residing in the western part of Copenhagen County was randomly drawn from the Danish Population Registry, in which all persons living in Denmark are registered (Andersen* et al.*, 1996).*

In Example 10.8, eight sex and age strata were formed (1, males born 1922; 2, males born 1932; 3, males born 1942; 4, males born 1952; 5, females born 1922; 6, females born 1932; 7, females born 1942; 8, females born 1952), and a random sample selected within each stratum.

There are many situations where this type of sampling is the most appropriate. Sometimes we may wish to have independent results for different strata, and to ensure an adequate sample size in each one. Also, there may be indications that prevalence will vary between strata (as in Example 10.8). Since different sampling schemes can be used in different strata, stratified random sampling is particularly convenient when, for instance, sampling frames are available only for some subgroups of the population (as in Example 10.9).

Example 10.9. *The World Fertility Surveys mentioned in Examples 10.1 and 10.2 used a general sampling design to select samples in the various participating countries. Firstly, a sampling frame of geographical area units whose boundaries were reasonably well defined was identified in each country. Secondly, a sample of these area units was randomly selected with probability proportional to the size of their population. Thirdly, a list of dwellings or households in the selected area units was drawn up. Fourthly, a similar number of dwellings or households was randomly chosen from each selected area unit and in these households all women meeting the criteria for entry into the survey were interviewed. Stratification was also used in some countries to ensure that both urban and rural areas were properly represented or because of the need to use different sampling designs in different geographical areas. For instance, population lists were available for some urban areas from which women could be randomly selected without the need to draw up a list of dwellings or households (Scott & Harpham, 1987).*

10.1.6 Study participants

Not everyone in the selected random sample will end up participating in the study. Some subjects will refuse to participate despite all reasonable efforts; others will have died or moved out of the area. Thus, participants are usually a subset of the initial random sample (as in Example 10.10).

All possible efforts should be made to ensure a high level of response and participation to minimize selection bias. The number of people or households interviewed, not just the number in the original sample, should always be reported so that non-response levels can be computed. What response level should be considered as acceptable in a survey? For an uncommon condition, a response rate of 85% might be unacceptable, because a few cases in the unexamined 15% might greatly alter the findings; on the other hand, in a survey of a relatively common attribute, this response level might be considered good.

Participants in any survey are likely to differ in some of their characteristics from those who do not respond. The important issues are whether this will introduce bias into the study and if such bias exists, how much it is likely to affect the results. In order to assess the bias introduced by non-response, it is essential to try to obtain some information

***Example 10.10.** In the cross-sectional survey described in Example 10.6, a random sample of women was selected in each of two geographical areas to assess (and compare) their prevalence of infections by human papillomavirus (HPV) and herpes simplex virus (HSV). The researchers described the method of enrolment of the study participants in each area as follows:*

"Eight hundred women aged 20–39 years, born in Greenland and residing in the municipality of Nuuk/Godthåb, were drawn at random from the computerized Central Population Register for the Danish Kingdom as a whole. Of these women, 104 had moved out of the municipality and one had died before they could be contacted, leaving 695 eligible for our study. […] General information about the study was provided in local news media (newspaper, radio), after which each [randomly selected] woman was invited, by a personal letter, for a visit to the local health clinic. Reminders were sent 2–3 weeks later and non-responders were finally contacted by a personal messenger from the clinic. Of the 695 eligible women, 586 (84.3%) were included in the study; 93 (13.4%) could not be reached and 16 (2.3%) did not want to participate. The relatively high proportion of women that could not be traced may be attributable to errors in the municipal population register and weaknesses of the postal service.

[...] From the Central Population Register a random sample of 800 women, born in Denmark, was drawn from the female population aged 20–39 years in Nykøbing Falster municipality. Fourteen of these women had moved out of the municipality and one had died prior to enrolment, leaving 785 women eligible for investigation. General information about the study was provided through local and national news media. Each [randomly selected] woman was then invited by personal letter to participate in the study and scheduled for a visit to the local hospital. Reminders were sent 2–3 weeks later and non-responders were finally approached by telephone. A total of 661 women (84.2%) were enrolled; 58 (7.4%) could not be contacted and 66 (8.4%) did not want to participate." (Kjaer et al.*, 1988)*

about the individuals who initially refused to participate or could not be contacted. Two approaches are possible. Firstly, a small random sample may be drawn from the non-respondents, and special efforts, including home visits, made to encourage their participation. The findings from this small random sample will indicate the extent of bias among non-respondents as a whole. Secondly, some information may be available for all persons listed in the study population; from this it will be possible to compare respondents and non-respondents with respect to basic characteristics such as age, sex, residence and socioeconomic status.

In Example 10.11, the age distribution of the participants was fairly similar to that of the total female population living in each of the two selected areas, with a slight under-representation of women in the youngest age-group.

Example 10.11. *Consider again the study described in Example 10.10. The age distributions of the total female population in the two selected geographical areas, of the two random samples selected from them, of the women who were eligible and of those who actually participated in the study are shown in Table 10.3.*

Table 10.3.
Age distribution of the female resident populations of Nykøbing Falster (Denmark) and Nuuk (Greenland), of the two random samples selected from them, and of the study participants.[a]

Age	Number (%) of women			
	Total female population	**Random sample**	**Eligible women**	**Participants**
Nykøbing Falster (Denmark)				
20–24	975 (27.9)	228 (28.5)	222 (28.3)	166 (25.1)
25–29	774 (22.2)	158 (19.8)	153 (19.5)	132 (20.0)
30–34	825 (23.6)	195 (24.4)	194 (24.7)	172 (26.0)
35–39	920 (26.3)	219 (27.3)	216 (27.5)	191 (28.9)
Total	**3494 (100)**	**800 (100)**	**785 (100)**	**661 (100)**
Nuuk (Greenland)				
20–24	582 (37.0)	281 (35.1)	227 (32.7)	193 (32.9)
25–29	439 (27.9)	226 (28.3)	192 (27.6)	171 (29.2)
30–34	328 (20.8)	167 (20.8)	156 (22.4)	127 (21.7)
35–39	225 (14.3)	126 (15.8)	120 (17.3)	95 (16.2)
Total	**1574 (100)**	**800 (100)**	**695 (100)**	**586 (100)**

[a] Data from Kjaer *et al.* (1988)

10.1.7 Final comments

Methods for selecting an appropriate sample constitute an important and well developed field of statistics that cannot be covered fully in this chapter. It is worth emphasizing, however, that the sampling design should be appropriate to the specific objectives of the study. To develop and implement a proper sampling scheme requires not only statistical expertise but also familiarity with the field conditions. For instance, if the survey uses 'households' as one of its sampling units, this term needs to be clearly defined in the context of the population where the survey is going to be carried out. A 'household' is usually defined as a group of people who live and eat together. However, in many societies, this definition will not be easily translated into practice. In such instances, it is crucial to obtain a good understanding and clear definition of the different living arrangements of the population to be surveyed before the study begins.

Other practical problems should be foreseen at the design stage of the study and unambiguous instructions written in the protocol and given to the interviewers. For instance, if there is no-one at home when the interviewer arrives, he/she should come back again rather than go to the house next door, because households with a person at home in the day-time tend to differ from those without.

Although a random method to select the study sample is generally the most appropriate, in a large number of epidemiological surveys it is not possible to select a sample in such a way for ethical or logistic reasons. For instance, studies that require the use of invasive diagnostic techniques that can only be performed in a hospital may have to rely on hospital attenders. The conclusions from such studies can be extrapolated to the population of 'hospital attenders', but the extent to which they can be generalized to the whole target population requires careful judgement.

The choice of an appropriate sampling design in cancer epidemiology will also depend on the aims of the survey. If the main objective is to obtain an overall estimate of the prevalence (or level) of an attribute in the target population, random sampling methods should be used at all stages of the sampling process to ensure that selection bias is not introduced. If, however, the main objective of the study is to examine potential exposure–outcome relationships, it may be more appropriate to select the main sampling units in a non-random way since, in these situations, informativeness is usually more important than representativeness. For instance, the selection of the 65 participating Chinese counties in Example 10.7 was not done randomly because the main objective of the survey was not to provide an overall estimate of the prevalence (or level) of the various lifestyle attributes for the whole of China. The intention was rather to compare the distribution of lifestyle attributes in counties known to have very different levels of mortality from certain cancer sites. It was, however, necessary to obtain unbiased lifestyle prevalence estimates for each of the selected counties and thus random methods were used to select the participants in each county.

Another issue that needs to be considered in the sampling design is the size of the sample. The sample should be sufficiently large to address the main objectives of the study with adequate precision, but not excessively larger than required, so that resources are not wasted. Sample size issues are discussed in Chapter 15.

10.2 Data collection

The methods used to collect the relevant data in cross-sectional studies are basically those discussed in Chapter 2. Questionnaires, sometimes supplemented by diagnostic tests and collection of biological samples, are most frequently used to obtain information from the study subjects in this type of study. Most questionnaires include questions about past exposures as well as current exposures. Information on past exposures considerably strengthens the ability of surveys to identify exposure–outcome relationships.

10.3 Analysis

Prevalence[a] is the measure of occurrence of a disease, condition or characteristic that can primarily be obtained from cross-sectional surveys (see Section 4.2.1). This is illustrated in Example 10.12.

[a] When the characteristic of interest is a quantitative variable (such as duration of breast-feeding, weight, height, etc.), prevalence can be calculated only if the observations are classified into categories. Otherwise, means or median levels can be used.

Example 10.12. *Suppose that a cross-sectional survey was carried out to assess the prevalence of breast cysts in a particular female population. A sample of 5891 women randomly selected from that population were examined and a total of 201 were found to have breast cysts. The prevalence of breast cysts in this population at the time of the survey was thus: 201 / 5891 = 3.4%.*

To examine the association between a putative risk factor for the attribute of interest and the attribute itself, the population is first subdivided into those exposed and those not exposed to the factor under study and the prevalence of the attribute in each group is calculated and compared. A *prevalence ratio* can then be computed as the ratio of the prevalence of the attribute of interest in those exposed to the putative risk factor relative to the prevalence in those unexposed.

Example 10.13. *Suppose that in the hypothetical survey described in the Example 10.12, the investigators wished to assess whether the prevalence of breast cysts was associated with having ever used oral contraceptives. The results are shown in Table 10.4.*

	Lifetime use of oral contraceptives		**Total**
Breast cysts	**Ever used**	**Never used**	
Yes	124	77	201
No	3123	2567	5690
Total	**3247**	**2644**	**5891**

Prevalence of breast cysts among ever-users = 124 / 3247 = 3.8%

Prevalence of breast cysts among never-users = 77 / 2644 = 2.9%

Prevalence ratio = 1.3

Table 10.4.
Breast cysts and lifetime use of oral contraceptives: data from a hypothetical cross-sectional survey.

In Example 10.13, the prevalence of breast cysts was 30% higher in ever-users of oral contraceptives compared to never-users. It should be noted that prevalence ratio is a good estimate of the incidence rate ratio only if the prevalence of the outcome of interest among those unexposed is low (less than 10%) and the duration of the disease is the same among those who were exposed and those who were unexposed to the factor of interest.

Most often the exposures we are interested in can be further classified into various levels of intensity as in Example 10.14. We can then examine trends in prevalence by level of exposure.

It is also usual to measure the strength of the association between a putative risk factor and the outcome of interest in a cross-sectional study by calculating the *odds ratio*. This is the ratio of the odds of exposure to a putative risk factor in subjects with the outcome of interest to that in subjects without the outcome. By calculating odds ratios, the cross-sectional study is analysed as if it were a case–control study. However, cross-sec-

Example 10.14. *A cross-sectional survey was carried out among women attending a university health service to investigate the determinants of cervical human papillomavirus (HPV) infection. A sample of 467 women were asked to complete a self-administered questionnaire on socio-demographic variables and sexual behaviour at the time of their visit to the clinic. A polymerase chain reaction DNA amplification method was used to detect HPV infection. The prevalence of HPV infection was then examined in relation to marital status and lifetime number of male sexual partners (Ley* et al.*, 1991). The results are shown in Table 10.5.*

Table 10.5.
Prevalence of HPV infection by marital status and lifetime number of male sexual partners.[a]

	No. of women	% positive for HPV	Prevalence ratio (95% CI)
Marital status			
Single[b]	437	47.4	1.0
Ever-married	30	20.0	0.4 (0.2–0.9)
Lifetime no. of sexual partners			
1[b]	90	21.1	1.0[c]
2–3	101	32.7	1.5 (0.9–2.4)
4–5	93	54.8	2.6 (1.7–4.0)
6–9	66	56.1	2.7 (1.7–4.3)
10+	102	68.6	3.3 (2.1–4.9)

[a] Data from Ley *et al.* (1991).

[b] Taken as the baseline category.

[c] χ^2 test for trend = 53.10, 1 d.f.; $P < 0.0001$.

(95% confidence intervals (CI) and χ^2 test for trend calculated using the formulae given in Appendix 6.1.)

The prevalence ratio for each exposure level was calculated by forming 2 × 2 tables as illustrated below for women with 10+ partners.

	Number of male sexual partners		Total
	10+	1	
HPV-positive	70	19	89
HPV-negative	32	71	103
Total	**102**	**90**	**192**

Prevalence among women with 10+ partners = 70/102 = 68.6%

Prevalence among women with one partner = 19/90 = 21.1%

Prevalence ratio = 68.6% / 21.1% = 3.3

tional studies differ from case–control studies in that the ‘cases’ and the ‘controls’ are defined *a posteriori*, i.e., during the analysis and not at the design stage. In fact, if the outcome of interest is quantitative, it is even possible to carry out several analyses using different definitions of ‘cases’ and ‘controls’ by changing the cut-off point.

Example 10.15. *In the hypothetical study of Example 10.13, we can calculate the odds of having ever used oral contraceptives among women with ('cases') and without ('controls') breast cysts.*

Odds of exposure to oral contraceptives among 'cases' = 124/77 = 1.61
Odds of exposure to oral contraceptives among 'controls' = 3123/2567 = 1.22
Odds ratio = 1.61 / 1.22 = 1.3

Note, however, that the odds ratio will yield a good estimate of the prevalence ratio only if the baseline prevalence of the condition is low (as in Example 10.15).

Example 10.16. *Using data from Table 10.5, we can calculate the odds of having had 10 or more partners ('exposure') among HPV-positive ('cases') and HPV-negative ('controls') women as:*

Odds of exposure among the cases = 70/19 = 3.68
Odds of exposure among the controls = 32/71 = 0.45
Odds ratio = 3.68/0.45 = 8.2

This odds ratio of 8.2 contrasts with the prevalence ratio of 3.3 calculated in Example 10.14.

In Example 10.16, it would be inappropriate to take the odds ratio as a measure of relative prevalence, because in this example the baseline prevalence of HPV infection is relatively high (21.1% in women who reported only one partner).

10.4 Interpretation

Cross-sectional surveys are relatively easy and economical to conduct and are particularly useful for investigating exposures that are fixed characteristics of individuals, such as ethnicity and blood group.

Cross-sectional surveys are not, however, the appropriate study design to investigate causal relationships because they are based on prevalent rather than incident cases. Studies of this type can reveal the presence or absence of a relationship between the study variables and prevalent (existing) cases. This implies a need for caution, since prevalent cases may not be representative of all cases of the disease. Cases of short duration, because of either rapid recovery or death, have a smaller chance of being detected in a one-time prevalence survey than cases of longer duration. It follows logically that cases of long duration are over-represented in a cross-sectional study. The characteristics of these long-duration cases may, on average, differ in a variety of ways from the characteristics of all cases of the disease being studied. Prevalent cases can also become unrepresentative of all cases when certain types of case leave the community. Some affected subjects may be institutionalized elsewhere or move to another city where there are special facilities for treatment.

Cross-sectional surveys are also an inadequate approach to the study of rare conditions, since it would be necessary to survey a very large population to identify enough cases. Thus, their use in cancer epidemiology has been limited to the investigation of factors associated with prevalence of precursor lesions.

Another major limitation of cross-sectional studies is their difficulty in establishing the time sequence of events. For instance, in our hypothetical example of breast cysts and oral contraceptive use, it cannot be assumed that oral contraceptive use preceded the appearance of cysts. In fact, women with benign breast disorders are sometimes prescribed oral contraceptives to improve their condition. In contrast, there would be no doubt about the time sequence of the cancer and such traits as blood type or maternal exposure to radiation.

Box 10.1. Key issues

- Cross-sectional surveys are studies in which a group of subjects (*sample*) is selected from a defined population (*source population*) and contacted at a single point in time. On the basis of the information obtained from the subjects at that point in time, they are then classified as having or not having the attribute of interest.

- Various methods may be used to select a representative sample from the source population. *Random sampling* is the best one, because it ensures that chance alone determines who is included in the sample, removing any possibility of *selection bias*. Different sampling designs may be used to select a random sample depending on the specific objectives of the study, availability of a suitable sampling frame, size and geographical spread of the source population, and costs.

- Selecting a random sample does not eliminate selection bias from the study. Selection bias may still be introduced into the study if those who participate differ in significant ways from those who refuse or cannot be traced. It is therefore important to ensure a high participation level.

- The main advantages of cross-sectional surveys are:

    1. They are easier to conduct than other individual-based studies because no follow-up is required.

    2. They provide a good picture of the health care needs of the population at a particular point in time.

    3. They can be used to investigate multiple exposures and multiple outcomes.

- The main disadvantages of cross-sectional surveys are:

    1. Being based on prevalent (existing) rather than incident (new) cases, they are of limited value to investigate etiological relationships.

    2. They are not useful to investigate rare diseases or diseases of short duration.

    3. They are not suitable to investigate rare exposures.

    4. It is difficult to establish the time sequence of events.

Further reading

* A practical book on the role, planning and conduct of surveys in developing countries is that by Casley & Lury (1987).

* Ross & Vaughan (1986) provide a methodological review of the use of cross-sectional surveys to obtain morbidity data and information on the use of health services in developing countries.