# EPIDEMIOLOGY

*A manual for distance learning*

WELLCOME TROPICAL INSTITUTE

This module was produced at the Wellcome Tropical Institute with the collaboration of Dr Roger Webber, London School of Hygiene and Tropical Medicine.

Other modules in this series:

- Acute Respiratory Infections
- Anaemia
- Burns
- Management
- Surgical Emergencies
- Obstructed Labour
- Sexually Transmitted Diseases
- Malaria
- The Red Eye and the Common Blinding Diseases

From 1986-1990 the Wellcome Tropical Institute, supported by the Wellcome Trust, worked with countries in Africa to develop the skills of rural and district medical officers. The Wellcome Trust is continuing to give limited support to work in distance learning. Comments and suggestions about this module and all enquiries will be welcomed by

Professor E H O Parry
c/o Department of Clinical Sciences
The London School of Hygiene and Tropical Medicine
Keppel Street
London WC1E 7HT
UK

## OBJECTIVES FOR THE MODULE

Study with this module will enable you to:

- Identify the sources of data about the community where you work, and its health, and assess their reliability, uses and limitations.
- Relate health and disease in individuals in a community to personal characteristics, place of residence and season of the year.
- Define and use measures of birth, death, morbidity and risk.
- Investigate an epidemic.
- Mount a programme of surveillance.
- Plan a simple epidemiological study in your district.
- Collect data for an epidemiological study in your district.
- Analyse data by sorting and summarising the information.
- Analyse data for causal associations.
- Assess the need for statistical analysis of data and perform specified statistical tests.
- Write a report of an epidemiological study in a clear, concise and logical manner.
- Critically evaluate the credibility and usefulness of an epidemiological report.

---

# CONTENTS OF THE MODULE

## REFERENCES PROVIDED WITH THE MODULE


1 Riley, L.W.; Waterman, S.H.; Faruque, A.S.G.; Huq, M.I. Breast-feeding children in the household as a risk factor for cholera in rural Bangladesh: an hypothesis. *Tropical and Geographical Medicine* (1987) **39**, 9-14.

2 Colley, J.R.T.; Holland, W.W.; Corkhill, R.T. Influence of passive smoking and parental phlegm on pneumonia and bronchitis in early childhood. *Lancet* (1974) **ii**, 1031-1034.

3 Swinscow, T.D.V. The $\chi^2$ tests. *British Medical Journal* (1976) **2**, 462-463, 513-514.

4 Swinscow, T.D.V. Correlation. *British Medical Journal* (1976) **2**, 680-681, 747-748, 802-803.

5 Newsome, D.A.; Milton, R.C.; Frederique, G. High prevalence of eye disease in a Haitian locale. *Journal of Tropical Medicine and Hygiene* (1983) **86**, 37-46.

# THIS MODULE AND ITS RELEVANCE

You learned epidemiology while you were studying medicine and you also had some practical experience during your posting in community health. You may well wonder what the purpose of this module is, particularly as it is described as a core module, a term which we use to describe a module which is central to your work and which can be used as reference and studied over a long period.

Just as all modules in your programme are designed to make you more effective in your work, so too is this module. We want you to be able to apply sound methods and do relevant studies, both in the community and in your whole approach to your work. We want you to find out through practical exercises and through reading what others have done, how exciting epidemiology can be. We want you to be able to tackle problems in your area and collect really useful data which will help to provide health care more efficiently (proper use of resources) and more effectively (objectives properly defined and reached by appropriate programmes).

This module is designed to be used as a resource when you study the other modules in this series. You will find that all the modules contain some epidemiology because it is so important and basic to our work. Thus the Obstructed Labour module considers women at risk, particularly in relation to the provision of maternal care and the demands of cultural practices. Mortality rates, seasonal variation and epidemics are included in the ARI module. The importance of morbidity, mortality and population structure on priorities in health service planning are dealt with in the Management module.

We hope you will refer to this module whenever you need help with epidemiology, either in your studies with other distance learning modules or in the course of your normal work. To find the information you need, you can either use the contents pages or the index at the end of the module.

One final point, do use all the tests and answer all the questions. You will find that they will make you think and use what you have learned. Only look at the answers provided when you have written down your own answers. You will be pleasantly surprised at how much more you will learn that way.

We hope you enjoy this module: we want you to find it interesting, as you work through it, and useful as you apply its lessons to your work. Why not begin your studies by answering the Pre-test on the next page to find out how much you already know and can do?

## PRE-TEST

*Question 1*

*You wish to study the health of the community in your district, using information that is readily available. List the different kinds of information that you will need.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Question 2*

*Following on from Question 1, list the places where you might find this information. How reliable is the information that is available to you?*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Question 5*

*You have received reports that there is serious malnutrition among young children in one part of your district. Describe the steps you would take in planning an epidemiological investigation into these reports.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Question 6*

*You want to collect data on infant mortality over the last two years in your district. Briefly outline how you will organise the collection of the data.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Question 3*

*When studying deaths in your district why is it important to calculate rates, in which you relate the number of deaths to the population?*

.....

.....

.....

.....

.....

.....

.....

*Question 4*

*List the different kinds of epidemiological measures you could use to study health and disease in the community where you work. Give the definition of each measure in your list.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

*Question 7*

*You have been given data on the height (in cm) of 100 boys and 100 girls aged five years in your district.*

1 *How would you present the data on height for each sex diagrammatically?*
2 *How could you summarise the data on height for boys and girls separately?*

.....

.....

.....

.....

*Question 8*

*Following on from Question 7, you have also been given data on the weight (in kg) of the same 100 boys and 100 girls. For the girls only, how could you examine the relationship between height and weight diagrammatically?*

.....

.....

*Question 9*

*Following on from Question 8, how could you determine whether, on average, boys weighed more than girls?*

.....

.....

.....

*Question 10*

*Why is it necessary to use standardised death rates when comparing different countries?*

.....

.....

.....

.....

.....

.....

.....

*Question 11*

*Outline, in not more than 100 words, the major elements that need to be examined when making a critical assessment of an epidemiological report.*

# INTRODUCTION TO EPIDEMIOLOGY

Much of your training has been in hospital where emphasis has been on the sick patient. You are taught how to take a history, examine the patient, do some investigations and come to a diagnosis. All these stages are necessary before you can formulate the correct course of treatment. When you are posted to a rural area you will find that you have so many sick patients that you are unable to give them all this kind of individual attention. However, when you look at the whole community you will see that there are health problems which are shared by everyone. For example, malaria might be affecting the whole community, whereas diarrhoea, measles and malnutrition are most prevalent among children. Many babies are being born, but few of these births are in hospital, so that your antenatal and postnatal care is limited to a small proportion of women only. How can you plan health services for these mothers? What is your strategy for vaccination for their children? How can you control the diseases common among the people? Epidemiology is the study of all these factors.

**The clinical history - a model for epidemiology**

Although occasionally we make a diagnosis without any history if the clinical picture is distinctive, as in severe pulmonary oedema or diabetic ketoacidosis, we would normally take a history from a patient and find out about his symptoms: what they are and how long he has had them. We would also ask him where he came from, what work he did and what his social and domestic habits were.

We must elicit the same kind of information in our epidemiological study. Collecting information is an important task of the doctor and can be absorbingly interesting and relevant, because the careful observer collects information in order to answer questions and to test an hypothesis which he or she has made, about a community's health.

Just as laboratory tests may be used in clinical diagnosis of an individual, to test an hypothesis about the cause of illness, so in a community diagnosis, data are collected about its health or disease, and are analysed by established methods.

Epidemiology is not just the study of diseases, it covers every aspect of health care. This might be the provision of health services, utilization of staff, even the deployment of limited resources into particular areas or at particular times of the year. One of the emergencies that not only sends panic through the community, but quite often the doctor as well, is an epidemic of a serious disease. The epidemiologist should be able to deal with all these, and we hope that this module will help you to manage them adequately too.

**Examples of epidemiological studies**

Some very interesting reports have come from medical officers in rural Africa and we have therefore given a few of these in this introduction so that you can see what can be done far from a well-equipped centre.

**'Tuberculosis'**

At most tuberculosis clinics there are patients who have curious shadows in their chest radiographs which do not disappear with treatment for tuberculosis. One such man in northern Nigeria[1] had had treatment for two years. He seemed fit but he was still coughing and the enigma of his radiograph continued. Only then was he asked what work he did and whether other men in the village had the same problem. He said that he cut grinding stones in deep and narrow pits which became very dusty and hot. His village was visited and a very high prevalence of silicosis (the stone had a high content of silica) was found among men, but no case was found in the women, nor in any men who had not cut stones.

As a result of looking at an individual in the context of his community, a group of men at high risk of a crippling disease was identified. The index case was used to launch a community study. The next step, obviously was to try to reduce the men's risk by finding ways of reducing their exposure to dust.

**Snake bite**

Another good example from northern Nigeria[2] concerns the epidemiology of snake bite. A doctor in an isolated mission hospital realised that the farmers in the savanna area served by the hospital were often bitten by snakes at the end of the dry season and beginning of the rains. They were clearing, hoeing and planting during those very months and so he collected data about admissions for snake bite month by month. (The snake, *Echis carinatus*, the carpet viper, is a terrible threat to savanna farmers). He showed clearly the seasonal fluctuations in the incidence, and mortality of *E. carinatus* bites. Obviously these seasonal data show that, in his management of the hospital *Echis* antivenom must be available abundantly during the peak season; the data also show that he must inform the primary care workers of the need to be vigilant during the peak season and to refer cases then. This study was very simple. It shows the need for a careful collection of facts, disciplined record keeping and the value of data collected over a long period.

---

[1]Warrell, D.A. *et al.* Silicosis among grindstone cutters in the north of Nigeria. *Thorax* (1975) **30**, 389-398.

[2]Warrell, D.A.; Arnett, C. The importance of bites by the saw-scaled or carpet viper (*Echis carinatus*): epidemiological studies in Nigeria and a review of th world literature. *Acta Tropica* (1976) **33**, 307-341.

**Burkitt's lymphoma**

You must remember the method used by Burkitt in his search for childhood cases of the tumour which now bears his name. He sent a questionnaire to hospitals in tropical Africa and he did a 10,000 mile safari to look for cases. Our mention of Burkitt's lymphoma is to record the careful seasonal and geographical (time/space) studies done at two mission hospitals and at one government hospital for five years from 1961 in the West Nile District, Uganda[3]. There was nothing spectacular about this work but it was careful, painstaking and continuous. The disease was shown, from the data, to have the epidemic characteristics of 'drift'. Patients whose dates of onset were close together tended to live closer together than could be expected on the basis of chance alone.

Why not think of similar studies in your district? How careful are your records?

**A definition of epidemiology and its range**

In the modules based on clinical problems we have often emphasised that you should approach a patient and consider him/her thus: 'Why should this person from this place fall ill in this way at this time?'

Think of this clinical approach as an example of the collection of data in a community - epidemiology is about information!

Epidemiology is the study of the distribution and determinants of health and disease in populations.This definition is on page 1 of *Epidemiology for the health officer: a field manual for the tropics*; edited by W.O. Phoon, published by WHO. What he writes in his introduction is so sensible that we have reproduced it in full here.

The denominator

*The most important idea in epidemiology is that of the denominator. It is vital to know about the apparently healthy non-patient as well as the obviously ill patient, and the population at risk of developing a disease problem as well as those who presently have it. The idea of the denominator is key to developing the needed information for health planning, supervision of activities, and the evaluation of health promotion and disease control activities. Every member of the health team should thoroughly understand that his responsibilities extend to those not-yet-sick as well as those already ill. The team must know how many children are to be immunised this month, how many patients are scheduled for leprosy treatment, how many houses are to have residual insecticides sprayed, etc.*

---

[3]Pike, M.C.; Williams, E.H.; Wright, B. Burkitt's tumour in the West Nile District of Uganda 1961-5 *British Medical Journal* (1967) **ii**, 395-399.

Factors which influence or cause a disease

*Epidemiologically, the idea of 'cause' is that of some factor which, when changed, results in a decrease in that particular disease. Although many interacting factors may combine to lead to a disease state, the Health Officer and his Team need to focus upon knowing which factors can most readily be changed to bring about a reduction of disease.*

*For example, the effects of malaria on the population were greatly reduced by the Greeks through draining swamps and building cities on hills. Protection from the bite of mosquitoes was found to be an effective preventive measure before the Plasmodium parasite was found. Today, a non-immune traveller visiting West Africa will develop falciparum malaria if a chemoprophylactic drug is not taken; for such a person, not to take an antimalarial is an effective cause of malaria. Thus, there are many levels of 'cause', and therefore various methods to reduce the impact of a disease on a community. The Health Officer/Team should concentrate on those methods which are most effective in their particular situation.*

*Rarely will the Health Officer or Team be involved in the type of epidemiological studies needed for unravelling the determinants of disease, but they must have the epidemiological knowledge and skills needed to acquire essential health information and to make the best use of it in the conduct of the health care activities under their responsibility.*

The collection of information

*Epidemiology is about information: the information needed for health planning, supervision, and evaluation of the health promotion and disease control activities. The key components of the data needed can be approached through a series of questions:*

*Who is affected, referring to age, sex, social class, ethnic group, occupation, heredity and personal habits? (These are the person-factors).*

*Where did it happen, in relation to place of residence, geographical distribution and place of exposure? (These are the place-factors).*

*When did it happen, in terms of the month, season or year? (These are the time-factors).*

*What is the disease or condition, its clinical manifestations and diagnosis?*

*How did the disease occur, in relation to the interplay of the specific agent, vector, source of infection, susceptible groups and other contributing factors?*

*Why did it occur, in terms of the reasons for the disease outbreak (e.g. breakdown of sanitary services)?*

*What now? The most important question - what action is now to be taken as a result of the information gained?*

The methods of epidemiology

*These include:*

- *Methods for the collection of health-related data, namely:*
  - *appropriate diagnostic tools for numerator data;*
  - *appropriate methods to estimate the population at risk for denominator data;*
  - *appropriate procedures for recording the data.*
- *Methods for the tabulation and analysis of the data to produce the essential information.*
- *The use of this information for decision making and evaluation.*

**The uses of epidemiology**

You may be able to make use of epidemiological information in your present post in the following ways.

- To make a community diagnosis, at a single time that is, to identify and to describe health problems in a community (this will usually be a cross-sectional study as observations are made at a specific time). Examples are the prevalence of anaemia or the weight for height profile of children.
- To monitor continuously over a period of time the change of health in a community. (This will be a longitudinal study as it involves the collection of information over a period of time and will commonly be related to measures which you have taken). Examples are the effect of a programme of vaccination, health education, nutritional supplementation, change of practices at work.
- To practise surveillance for specific diseases. Examples are cholera, in order to be able to act quickly and so cut short any outbreak, or pertussis, a useful monitor for the cover achieved by an EPI programme.
- To investigate an outbreak of a communicable disease, analyse the reasons for it, plan a feasible remedy and carry it out, and monitor the effects of the remedy on the outbreak.
- To plan effective health services. Note: this derives from the first use, community diagnosis. Effective services, interventions and remedies all depend on accurate community data. The responsible medical officer or health officer is thus able to do things which will really help the people among whom he lives and works.

---

# UNIT 1 HOW TO LOOK AT YOUR COMMUNITY: 1

# UNIT 1: OBJECTIVES

Study with this unit will enable you to:

- Identify the sources of data about the community where you work and its health, and assess their reliability, uses and limitations.
- Promote the collection of accurate data by members of your health team.
- Collect data about the characteristics of the population.
- Relate health and disease in individuals in a community to place of residence and season of year.
- Establish birth rates and death rates in your community.
- Calculate standardised death rates.
- Give reasons for each of your decisions or for the choices you make in all the objectives listed above.

---

# UNIT 1: CONTENTS

# INTRODUCTION

Community leaders

We assume that you have established good relationships with the leaders of the community and that your work is done with the knowledge and cooperation of the community leaders. If it is not, you will not get very far!

Health problems

You will have talked to the leaders and they will have told you of the particular health problems of their people and what they expect you to do about their health. Obviously you want to help them but you may wonder how to start. This unit will help you.

Information

If you are to provide effective health care for the community, you will need to have some information about the people and the environment they live in. Firstly, you should find out what data exist already and make use of this information. Then you will need to look at your community in order to answer certain questions about the problems which you have been told about, or have identified. You will need to find out:

Who?
- Who is affected?

Where?
- Where are they affected (where do the problems occur)?

When?
- When are they affected (when do the problems occur)?

By asking these questions you can identify those people who are at particular risk from a disease or health problem, and who will therefore benefit most from specific health care activities and programmes. For example, you may find that measles is an important problem in children in your district. But although it is present all year, most cases occur just before the rains start. It is therefore important that the measles immunisation programme is intensified two or three months before this peak period.

Targeting health care

Mortality rates

Finally we will consider how a study of birth and death rates can provide important information on the health care needs of the community. For example, indicators like specific mortality rates can be used to identify particular health care priorities, to monitor health care delivery and to evaluate the effectiveness of control or preventive programmes.

---

# 1.1 INFORMATION

Using routine data

The assessment of the health status in the community is the basis both for planning and evaluation of the health services. Useful information needed for making decisions can often be obtained from routinely available data, even though these are not accurate or complete enough for detailed or elaborate analysis. We shall consider in this section what information you can obtain on the frequency and distribution of morbidity, mortality and their causes from routine sources.

## SOURCES

*Activity*

*Get a member of your health team to make a list of the reports and data which you think might be of use to you and assemble them.*

Sources of information

Now that we realise that health is influenced by many factors, you will need to get your assistant to find out what has been written on agriculture, education, water supply, or any other subject which tells you about the people among whom you work. We have made two check-lists of the sources of information which may be available to you: one on health-related sources and the other on non-health related sources. Look carefully at these and, through the work of your team, assemble as much relevant information as possible.

**Health-related information sources**

Dispensary/aid post

- Dispensary and village aid post registers: a register is kept of all patients seen in every dispensary and, perhaps, in most aid posts also. This will contain details of name, age, sex, diagnosis and treatment, for the number of patients seen each day.

Health centre/hospital

- Health centre and hospital returns: these will either be outpatient or inpatient records. Outpatient records will be similar to those above. Inpatient records will have preliminary and final diagnosis. Cause of death, occasionally supported by post-mortem evidence, will be given.

MCH

- Maternal and child health: MCH might form part of a combined report from a dispensary, health centre or hospital or may be a separate report for all health services.

Laboratory records

- Laboratory records: although laboratory records do not cover all diseases, a common investigation such as a blood slide can give an idea of the proportion of parasite-positive malaria cases as opposed to those with 'clinical' malaria. Identification of malarial parasite species or of schistosome species is important in planning control programmes. For certain communicable diseases, such as cholera, (where bacteriological confirmation is the only proof of diagnosis) the laboratory records are the most reliable source of information about the incidence of the disease (see Unit 2).

Campaign reports

- EPI or other campaign reports: these often have their own census data.

Notifiable disease registers

- Notifiable disease registers: notification systems are restricted to selected 'important diseases', usually infectious ones which require prompt action for control.

**Non-health related information sources**

- Census.
- Survey - agricultural
  - economic
  - social.
- School - enrolment and class lists.
- Factory accident records.
- Plantation or factory registers of workers.
- Maps/mapping survey.

**USES AND LIMITATIONS**

Now let us return to the sources of data from units of the health service in your area.

*Exercise 1*

*What do you think are the uses and limitations of the three health-related sources of information shown in the table on the next page? Think in terms of the people served and the level of health care provided and then write down your ideas in the spaces provided in the table. When you have done this compare your answers with ours on the following page.*

For example, a nutrition clinic would help you to find cases of marasmus, but would give you no idea about the general health of children in the area served by the clinic because mothers choose to take to the clinic only those of their children about whom they are worried.

| Uses | Limitations |
|---|---|
| Dispensary and village aid post registers | |
| Health centre and hospital returns | |
| Maternal and child health | |

Dispensary/aid post

Dispensary and village aid post registers are useful for finding out overall numbers of people presenting for treatment and could be a source of data on births and deaths.

Births and deaths

Clinical diagnosis is generally of little value. In a survey that I conducted in one country I found that it was not worth asking for the completion of a disease list of some 20 complaints as only data on diarrhoea, measles, chicken pox and new epidemic diseases were reasonably filled in, even in the field by trained health workers!

Health centre/hospital

Health centre and hospital returns are likely to be accurate with respect to disease diagnosis but the data may only relate to the area served by the hospital. Time-based data, such as length of stay, and organizational information, such as staffing or the distance patients travel to hospital, can be used also.

MCH

Maternal and child health clinics are a valuable source of information for comparing antenatal care, deliveries and the subsequent fate of the children.

Immunisation

If you compare the number of births with the number of children immunised, this can give an indication of the coverage of any immunisation programme.

Childhood diseases

MCH clinics are one of the best sources of data on childhood diseases such as measles and malnutrition and, over a period of months or years, are reasonably accurate. MCH records, alone, are not enough as they are only a source of data on births and on deaths in children under five years. Use other sources of data to obtain a more representative picture. MCH records can also be used to measure the workload of the MCH workers.

**Inpatient and outpatient records**

Hospital inpatients

Analysis of hospital records can provide high quality information on the most important causes of major illness in a community. But to be useful as an indicator of the health status of the population you must make allowances for the fact that patients treated in hospital are not representative of the general population in the area. People from remote areas, infants and the elderly, for example, will be under-represented. In some countries, many if not most, seriously ill patients never reach hospital.

Outpatients

Records of outpatients seen in hospitals, health centres, health posts and clinics often provide much ill-defined data. Diagnostic data are usually given in terms of the chief complaint. Those coming for immunisations or other preventive services may be included with those who come because of illness. The patients who are seen are again probably not representative of the general population: although coverage of the population may be greater than with a hospital because of greater geographical distribution, the people who live near a facility or who can afford the time to come will be over-represented. However, these records do provide information about the usage of outpatient facilities and the most frequent complaints, and may help you to understand the pattern of disease in your community.

**Problems with routine data**

The routine data sources which we have examined so far in this unit will almost certainly fail to include a great deal of important illness and disability. In particular, much of the chronic illness due to tropical diseases such as schistosomiasis and leprosy, blindness, under-nutrition and crippling due to birth trauma or polio, will not be detected from routine records. If you want to get information about these kinds of condition you will need to carry out an epidemiological survey (see Unit 3).

Chronic conditions excluded

Another problem with most routine data is that they relate only to numerator data. The usefulness of expressing data in terms of rates, for which denominators are needed is discussed in Unit 2.

Numerator data

**ACCURACY**

You may well feel rather sceptical about the accuracy of some of these data and it is indeed true that reports may have data which were not checked: a systematic error may be present or other data may have been entered which were not true.

For example, data about the age of individuals could be inaccurate for several reasons:

Sources of inaccuracy about age

- Older age is highly respected - people will add to their age!
- Where there is no tradition for counting age by years, events have to be used. For adults and children, therefore, the date of their birth, or their marriage, or their first child's birth has to be related to an event.
- An inaccurate observer may round off an age to the nearest five years or may routinely suggest '40 years' or '25 years'! [When we looked at the ages given in the registration records of a major teaching hospital, the registration clerks classified adults in just this way and so the ages which they recorded were of very little value].

It is very important to assess the accuracy, and check the reliability, of any reports which you have if this is possible. Reliability is dealt with in more detail in Unit 3.

Made-up figures

When someone makes up data about disease in a population he tends to repeat the same figure as was recorded in the previous month or round it off to an approximate one hundred or so. If figures fluctuate within reasonable limits, then suddenly become twice or half what they were, you should investigate the source of information. Disease patterns show trends, either up, down, or they remain the same. It would not be reasonable if the figures suddenly doubled or trebled in one month and then returned to normal in the following month, unless there was an epidemic!

Increasing the reliability of data

There are several ways by which you could make the data for your area more reliable. Here are our suggestions.

- You could train all the members of your health team to collect accurate data, to avoid bias and to record carefully, and you would check the accuracy of their work. (See Units 3 and 4).

Different sources

- You could take the information from a number of different sources. If you then compared the data from the different sources you might well be able to identify inconsistencies and thus inaccuracies.

Practical suggestions

Few data are wholly accurate. What degree of inaccuracy can you tolerate? This cannot be expressed in figures. The best you can do is summarised as follows.

- Be clear about the data you really want.
- Decide how best this can be collected most accurately within the limits of your resources.
- Explain the reasons and methods carefully to the members of your team.
- Check from time to time on the data which are being collected.
- Let your colleagues and field workers know how the data they have collected has helped your work.

Do not wait for ideal data but rather demonstrate how data can be used and, in so doing, make the case for better data to be collected. On the other hand, it is a waste of time to continue collecting useless and unnecessarily complicated data.

---

## 1.2 CHARACTERISTICS TO BE RECORDED

The distribution of a disease is described in terms of the personal characteristics of those affected, for example their age and sex, together with the variations in disease occurrence at different places and at different times. In this section we will consider some of the attributes and variables which are basic to the epidemiological distribution of a wide range of diseases.

Definitions

An attribute is a quality or characteristic of a person, such as sex or cultural group.

A variable is a quantity which may vary in value. Examples are age, height or blood pressure.

### PEOPLE

You will have learned the reasons for collecting facts about people when taking histories of patients. Exactly the same principles apply when you want to collect facts in the community in order to help you make decisions about its health. For example, you noticed that apparently nearly all the men with cough seem to come from a particular village. You therefore collect data about all the men you see, particularly noting their ages, occupations, where they live and where they work.

Population characteristics

The distribution of a disease is described in terms of the personal characteristics of those affected. For example:

- Age.
- Sex.
- Marital status.
- Ethnic group.
- Social and economic standing.
- Occupation
- Religion. (Beliefs affecting eating or other habits may influence health).

The following examples concerning social/economic groupings and occupation illustrate the importance of considering these factors when studying health and disease in a community.

Social/economic groups

In the industrialised countries five social/economic groups are recognised. It is far less easy to draw these divisions in Africa. Instead you will have to follow a pattern which is applicable to the area where you work. This may be quite different and more limited than what pertains in the capital city.

If you plan a project which demands social/economic data, discuss any classification which you propose with your supervisor first.

Occupation

Let us briefly consider the relationship between occupation and health or disease. If those whom you survey have jobs, find out how long each individual has been in a job. Do not accept the name of the job. Make certain that you define what is done and whether this exposes the individual to any hazards.

Occupation can influence health/disease in various ways, for example:

- In some occupations people are exposed to specific agents of disease, risks or toxins - for example lead (car battery workers), dusts (miners), alcohol (bar men and women).
- Some occupations select specific types of individual.
- Occupation may govern, or be associated with, social and economic standing. For example, unemployment is an important factor in leading to sub-optimal health.

The seven characteristics we listed on the previous page are the basic facts which you have to collect about any person in a community in order to give you a baseline from which you can work, or, put in another way, a profile of the community. You will then be able to identify how any particular group of people differs from the normal values you have established. Additional data on other characteristics will have to be collected and analysed when specific problems are to be studied.

**PLACE**

The place where an individual lives can profoundly affect health. For example, if people have moved from a different area to the place where they now live they may not reveal that they are migrants, so that vital clues about disease may be lost unless you find out for how long they have lived in that place.

Maps

This is where study of the geography of disease begins. If you find epidemiology rather dull, try maps and medicine, or maps and disease patterns and I am certain that you will become fascinated and will want to know more. Here are some important points to remember in relation to disease distribution.

Methods

- Record accurately where an individual lives.
- Make certain that the place is defined - house, (street), village, district, so that it can be found if follow-up is needed.
- Do not be content with the address given on a hospital admission card.
- Make certain that those members of your team who collect geographical data do so very carefully.
- Assemble relevant maps in your office.

Mistakes

An individual may give his address as the place where he spent the last night, or he may name his birthplace although he left the place in infancy.

Migrants

- Urban migrants
- Pastoral nomads. Try to find out the area where they graze their herds both in the dry and in the rainy seasons.

Micro and macro

- Distribution of disease may vary over small areas in the same climatic zone if trees, standing water or a river produce a small micro-climate (ideal perhaps for survival of tsetse fly).
- Geographical features are important - temperature, rainfall, altitude and type of soil.

We now give a few examples of the usefulness and interest of studies of the geography of disease.

Examples of the importance of place

Elephantiasis was found to be common in parts of Ethiopia and neighbouring Rwanda and yet these were areas where bancroftian filariasis was unknown. The soil in these areas was rich in silica: the tissue response to the silica was the probable reason for the blocked lymphatics.

The position in a ward or room (and procedure undergone by patient) are all relevant in locating cross-infection.

Accurate mapping of houses and the origin of victims may quickly indicate a common source of a pathogen in water or food.

Plotting of villages affected or of compounds in villages, can determine a policy to contain the infection. For example, in an outbreak of meningococcal meningitis, selective vaccination to susceptible people in compounds from which index cases came reduced secondary cases.[1]

Performance by health care teams can be evaluated if different districts are compared.

'Macro' data within a country may give clues to the need for remedial action. However many geographical differences are not understood: for example the prevalence of epilepsy in some areas of Tanzania.

**TIME**

From place, we now move to time, which is most obviously, but is by no means only, concerned with seasonal changes.

Seasons

In tropical Africa the seasons of the year are very important to people who live close to the land and who do not have modern energy-consuming means which can raise or lower ambient temperature or pump water at all times of the year.

When we study the health of a community we think over long periods whereas when an individual is studied acute events are often dominant.

Asking questions

Try to get into the habit of asking why should this happen to this community at this time or when did the problem arise? Do not be dismayed if you are unable to explain what you observe: the first essential is to be alert and to observe; the second essential is to

---

[1]Greenwood, B.M.; Wali, S.S. Control of meningococcal infection in the African meningitis belt by selective vaccination. *Lancet* (1980) **i**, 729-732.

record; the third is that you should follow classical scientific method by formulating an hypothesis (see Unit 3) to explain what you have observed and recorded, and ideally you should test your hypothesis.

Hypotheses

Whether you are right or wrong is less important than your collecting data and acting on it scientifically, to improve the health of people in your area.

The subject of time can be considered further in three ways: epidemics, cyclical changes and secular changes.

Epidemics

Any notable increase in the incidence or prevalence of a disease is an epidemic. (For a detailed account of epidemics see Unit 2). In the Savanna belt of Africa meningococcal meningitis is a classic example. Another example is the epidemics of cholera in Africa since 1970.

Cyclical changes

In Africa, cyclical changes in disease are seasonal. Be alert for them as they could be very important in your area.

Secular changes

This term is used for changes over a long period and may be a result of:

- Changes in the age structure of the population and in its immunity;
- Changes in pattern of life, for example the rapid urbanisation in Africa;
- Changes in the accuracy of diagnosis or in reporting of a disease.

## 1.3 BIRTHS AND DEATHS

Recording systems

Information about birth and death is shared with other disciplines and forms the basis for the discipline of demography. In rural Africa most births and deaths are not witnessed by members of the health team so that additional methods are needed to collect data about them. A national registration system is unlikely to have been developed in most countries, but village or community leaders are in a unique position to fill the gap. They may already have such a system: if not, together with their teams, they can be asked to start to keep simple records. If it is not possible to obtain information this way then it might be necessary to do a sample survey (see Unit 3) or even a census.

### BIRTH

Required data

Information required on birth will be date, sex, place, name of mother, name of father, type of delivery and weight of infant. Ideally, if the data are needed for planning maternity services, find out about birth order, birth interval, number of children dying, age of mother's marriage, number of co-wives and religion.

Uses of birth data

Birth data are used for the calculation of perinatal mortality rates (see later) and for working out the increase in population. Population figures are required in planning health services as the rate of increase indicates the degree of staff expansion and facilities.

### DEATH

Unreliability

Mortality data are highly unreliable in much of Africa because most people want to die at home, necropsies are few, and medical attendants are not present at death. The date, age, sex, place and cause of death are needed, but it is particularly difficult to be accurate about the cause of death unless a knowledgeable medical officer, assistant medical officer or health officer is present.

Cause of death

The cause of death has to be subdivided into the immediate cause of death and the conditions underlying the cause. For example, the cause of death may have been pneumonia, while the underlying illness was kwashiorkor. A WHO committee considering this problem has recommended a very useful list of causes of death which can be used by people who know little about medicine. It is shown in Figure 1, modified slightly from the original and with suggested diagnoses in brackets.

Information on death is set against births in calculating the rate of population increase. Death rates could be important to you as a district medical officer or as a health planner.

Figure 1

NOTIFICATION OF DEATH
(Not occurring in hospital)

Name..... Date of death ..... Sex ..... S.M.W.D.

Address:..... Place of death: Village.....

..... Province.....

Birthdate or estimate of age:..... Occupation:..... Race:.....

Circle below the condition which most closely describes the cause of death.

1 Diarrhoea.
2 Cough for more than three months with or without blood in sputum and loss of weight (tuberculosis).
3 Cough of short duration with high fever, shortness of breath (pneumonia).
4 Intermittent high fever, rigours (malaria).
5 Rigid neck, fever of short duration, headache (meningitis).
6 Fever with rash (measles, chicken pox or specify).
7 Lock-jaw, spasm of the muscle, history of wound and/or childbirth (tetanus).
8 Sudden death including stroke (coronary thrombosis or C.V.A.).
9 Increasing breathlessness, swelling of ankles and/or abdomen (cardiac failure).
10 Chronic cough, breathlessness, asthma (bronchitis).
11 Acute abdominal pain, abdominal rigidity (peritonitis).
12 Complete stoppage of urination (renal failure).
13 Abortion.
14 Other complications of pregnancy (specify).
15 Complications of delivery (specify).
16 Complications of puerperium (specify).
17 Death of the newborn within seven days.
18 Malnutrition.
19 Transport accidents.
20 Accidental poisoning.
21 Bites/stings of venomous or other animal (specify animal).
22 Falls.
23 Burns.
24 Suicides/homicides.
25 Drowning.
26 Other injuries and accidents (specify).
27 Senility (old age).
28 Unknown causes.
29 Other causes of death (give full details of symptoms duration and possible cause).

Remarks:

Date: ..... Signature:.....
Name and address of reporter

Death rates and planning

Death rates vary by age groups, but you should look carefully at the data in your district to see whether there is an unusually high rate in any age group. If the figures you have are reliable, you can use the death rates as a basis for planning a more effective health service to the vulnerable groups. Then, over years, the rates should show how successful your work has been.

**DEATH RATES**

Death rates are described as crude, when they refer to the entire population, or specific when they refer to a particular section of the population according to age, sex, occupational group, or other variable. Death rates are normally recorded as annual death rates. Since deaths are occurring throughout the year, the mean of the population, or the mid-year population, is used as the denominator. The crude death rate is calculated thus:

Crude death rate

$$\frac{\text{Number of deaths in a year}}{\text{Mid-year population}} \times 1000$$

Note that rates are conventionally expressed to the base 1000, that is they are to be multiplied by 1000 and the result is then stated as so many 'per 1000'.

Uses

The crude death rate therefore measures the proportion of the population dying every year or the number of deaths in the community per 1000 population. It reflects not only mortality risks but also the population composition, as the death rate varies with the population structure just as the disease rate does. Therefore you can only use the crude death rate to compare relative mortality in two populations if they have a similar age/sex composition. If you wish to compare the mortality rates of populations with different

Standardised death rate

compositions, you need to calculate a standardised death rate. An age-standardised rate is used to eliminate the effect of age differences in the populations being compared; an age/sex standardised rate is used to eliminate the effect of differences in the age and sex distributions. Refer to page 39 for information on the standardisation of death rates.

Standard populations

With standardisation, different results will be obtained when different standard populations are used. If two populations are compared directly this does not matter, but where a number of different populations are to be compared (e.g. of several countries) or the rates are to be compared over a period of time (changes from year to year) it is necessary to have a defined, unchanging standard population. This has to be agreed but, once chosen, it must be used consistently. One method is to select a particular year and compare all other death rates with that year. This gives the standardised death rate for that particular country.

Comparative mortality index

A derivation of this method is to use the comparative mortality index where age and sex mortality of the current year are compared with the standard year, and an increase or decrease noted. This can also usefully be applied to occupational groups.

## SPECIFIC MORTALITY RATES

This is revision for you, but it will help you when you plan your services, teach your team, or even perhaps prepare yourself for an examination. The mere repetition of a series of mortality rates is meaningless. But it is highly relevant to calculate specific mortality rates as they will help you identify vulnerable groups for whom you can then plan improved services. Think therefore of these rates as a means to the end of better health care!

Age-specific death rates

Because of the profound effect of age on mortality, it is necessary to construct death rates for each age group and to use these rates for comparison.

Sex-specific death rates

Death rates can be calculated separately for males and females to give sex-specific death rates.

Uses

As specific death rates do not summarise the total mortality, comparison of overall mortality conditions in two populations is cumbersome because of the need to compare rates for all the different age groups and for males and females. Age/sex standardised death rates should be used instead. However, specific death rates are used to measure the risk in specific age and sex groups and give the essential components for constructing life-tables. Let us review some other important specific mortality rates.

Stillbirth rate

A stillbirth describes the birth of an infant after the 28th week of pregnancy, which did not, after delivery, breathe or show any other signs of life. It is calculated thus:

$$\text{Stillbirth rate} = \frac{\text{Number of stillbirths}}{\text{Total of births, live and still}} \times 1000$$

If the rate is high in your area, the quality of antenatal care and the conduct of childbirth must be suspect.

Perinatal mortality rate

Similarly the perinatal mortality rate reflects the antenatal care and management of delivery. It is calculated thus:

$$\frac{\text{No. of stillbirths + deaths between birth and 7th day}}{\text{Total births, live and still}} \times 1000$$

Obviously if you find a high perinatal mortality rate you will have to get out and find out why. Are antenatal clinics poorly supervised?

Neonatal mortality rate

Neonatal mortality is largely due to prematurity, malformations, accidents or injuries at birth, and lack of cleanliness and sterility during or after delivery. In addition it reflects the adequacy of antenatal care. It is calculated thus:

$$\frac{\text{No. of deaths of infants under 28 days}}{\text{No. of live births}} \times 1000$$

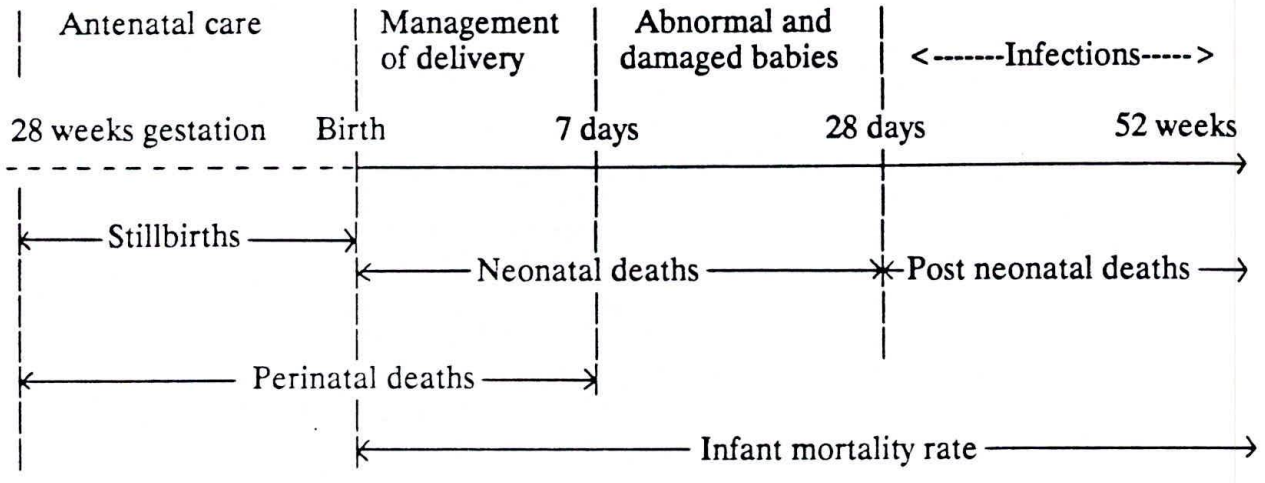
Infant mortality rate

The infant mortality rate, after the first 28 days, reflects the health of the community in which the child is being brought up. Thus it is high among people who have little health care, chiefly because infections, such as ARI, diarrhoea and malaria, are common among their infants. It is calculated thus:

$$\frac{\text{No. of deaths of infants under 1 year of age}}{\text{No. of live births occurring during year}} \times 1000$$

These different rates are illustrated in Figure 2.

Figure 2 Childhood mortality rates




**Maternal mortality rate**

The maternal mortality rate is the number of maternal deaths ascribed to puerperal causes per 1000 total births. It reflects the standards of all aspects of maternal care (antenatal, delivery and post-natal). It is calculated thus:

$$\frac{\text{No. maternal deaths due to pregnancy, childbirth \& puerperium}}{\text{Total births live and still}} \times 1000$$

**Cause-specific death rate**

Death rates for any specific disease, such as pneumonia, may be stated for the entire population, or for any age or sex sub-group. The cause-specific death rate is useful for indicating major causes of mortality in the population. It is calculated thus:

$$\frac{\text{No. deaths due to a specified disease in a year}}{\text{Mid-year population}} \times 1000$$

**Case fatality rate**

The case fatality rate represents the probability of death among diagnosed cases or the killing power of a disease. It is typically used in acute infectious diseases such as ARI. Its usefulness for chronic diseases (even when they are infectious) is limited because the period from onset to death is typically long and variable. The case fatality rate for the same disease may vary in different epidemics as the balance between agent, host and environment alters. It is calculated thus:

$$\frac{\text{No. deaths due to the disease in a specified period}}{\text{No. of cases of the disease in the same period}} \times 100$$

## BIRTH RATES

The annual birth rate is a 'crude' measure of all the influences which determine the rate at which a population reproduces itself. It is calculated thus:

Annual birth rate

$$\frac{\text{No. of live births in a year}}{\text{Mid-year population}} \times 1000$$

When you look at data about birth rate, make certain that you know the source(s) of the data. If they refer to a place where health care is effective, as in a maternity hospital, the rate will be falsely inflated. Births must be corrected to normal place of residence of mother. If they are to be compared, then they must be standardised in the same way as death rates.

The birth rate is not strictly a measure of the probability of birth among the 'susceptible population at risk', i.e. females of child-bearing ages. In its present form the birth rate under-estimates the probability of birth because the denominator includes sections of the population not able to give birth (males, old people, children).

Fertility rate

If we want to measure fertility, we use the fertility rate which takes into account the number of women of child-bearing age. It is calculated thus:

$$\frac{\text{No. of live births in a year}}{\text{No. of women aged 15-45 years in the population}} \times 1000$$

Rate of natural increase

The rate at which a population is increasing is measured by the rate of natural increase and is simply calculated by subtracting the crude death rate from the crude birth rate, thus:

Crude birth rate - crude death rate

From the rate of natural increase, the future population can be forecast. This is further calculated as for compound interest, so a rate of natural increase of 3% means a population will double itself within about 23 years.

Migration

Note that this assessment of population increase does not take into account migration into and out of the community. A truer measure of population growth would be:

$$\frac{(\text{Births} + \text{in migration}) - (\text{Deaths} + \text{outmigration})}{\text{Mid-year population}}$$

## STANDARDISATION OF DEATH RATES

Crude rate

Because the overall mortality pattern of a population depends on its age and sex composition, the crude death rate is a poor comparative index of total mortality. If we want to compare the mortality rates of different populations, we therefore need a summary index of total mortality that is not affected by the differences in age and sex composition of the population. This

Standardised rate

summary rate is called a standardised death rate. To eliminate the effect of age differences in the populations being compared, we use an age-standardised rate; to eliminate the effect of differences in both age and sex distributions, we use an age/sex-standardised rate.

Standard population

In concept, the age-standardised death rate of a population is the overall death rate that the population would have if it had a

# UNIT 2: HOW TO LOOK AT YOUR COMMUNITY: 2

## UNIT 2: OBJECTIVES

Study with this unit will enable you to:

- Count events related to health and disease, and describe them by appropriate terms.
- Estimate the degree of risk related to a factor which affects the health of the community or of a group of individuals.
- Investigate an epidemic, and identify why it is spreading and how it can be stopped.
- Mount a programme of surveillance, either as an emergency or as a routine.

---

# UNIT 2: CONTENTS

# INTRODUCTION

In the first unit we looked at the many variables which we might consider when we make a community diagnosis. But you are a health worker/medical officer because communities are not just composed of people, but of people who may be ill now or vulnerable to disease later.

Measuring disease

Firstly, therefore in this unit we shall look at the measurement of disease in a community - its incidence and its prevalence. These measurements will help us to define the important disorders which affect health. Perhaps you have noticed that a certain disorder is common in the district where you are working, whereas it was rare in the teaching hospital where you trained.

Measuring risk

We will then consider the measurements we can use to identify those at risk of disease so that we can improve health care for these people through more effective planning and health education.

Epidemics

Epidemics of infectious disease, such as cholera, typhoid and meningococcal meningitis are an important threat in many tropical countries. Modern epidemiology arose out the of the study of such epidemics: outbreaks of disease affecting many people and caused by infections, or in some cases dietary deficiency or poisoning. The third section of this unit deals with the investigation of epidemics so that you will be able to act promptly and effectively.

Surveillance

Effective surveillance systems are essential for the control and prevention of disease. The final section of this unit deals with surveillance: what it means, what it is used for, and how it can be done.

---

## 2.1 COUNTING DISEASES

### NUMBERS AND RATES

Let us revise the terms used for basic epidemiological measurements.

Number of cases

The number of cases is the simplest of all measurements and describes the number of people with a defined disease in a defined population. For example: 'There were 33 cases of snake bite among 1295 adult male admissions to a district hospital in 1976', or, 'There were 35 cases of rabies in Ghana - population 15 million'. This merely tells us that a disease exists and gives an idea of the burden on the health services of a country. More accurately we express disease frequency as rates: incidence and prevalence.

Rates

If you are looking at trends (general course of events) over a period of time, or comparing the risks of disease occurrence between communities or particular groups, you may not be able to draw valid conclusions if you record only the number of cases. The population size of each group or community to be compared must also be considered. In this case the information should be expressed in terms of rates. When we calculate the rate of occurrence, we are relating events or cases to the population which has given rise to the cases.

Population at risk

This 'population at risk' refers to the risk group of people who have the potential of getting the disease and this may contribute to the number of cases. Such a 'population' may refer to the whole population of a country or to sections of it.

A rate therefore has the following general formula:

Numerator: Number of cases in a specified time period
Denominator: Population at risk in same time period

It is usually expressed as per 100, per 1000 or per 100,000 depending on the actual figures obtained. Do not choose a constant which results in a large number of figures or decimal points (see below).

### INCIDENCE

Definition

You can remember what this means from the word incident: an event. Incidence refers to the number of **new cases** (or events) of the disease which occur during a specified period.

*Exercise 1*

***For example: in an area of a country with an estimated mid-year population of 500,000, 40 new cases of kala-azar were reported in that year. What is the incidence rate? Relate this to every 100, 1000 and 100,000 people.*** [The answer is on page 50].

.....

.....

**Uses**

Incidence rates can be used to make statements about the probability or risk of disease. They are compared among population groups with different exposures or attributes in order to measure the influence such factors may have on the occurrence of disease. Thus, estimates of relative risk are obtained (see section 2.2). When incidence rates

**Standardisation**

are used to compare the frequency of disease in different populations, it is necessary to standardise the rates so that a different age/sex structure in each population does not distort them (see page 52).

**Example**

In your district you can determine the incidence of a disorder by starting a longitudinal study (see Unit 3) in which you identify every new case of, say, tetanus in the district. Do all cases of tetanus come to hospital? They probably do. If so, and if you know the population, you can calculate the incidence rate accurately. Also if you find that the disease has its highest incidence in neonates, you will immediately be able to act as a result of the data you have collected.

**Tetanus**

If neonates are developing tetanus, it is most probable that many deliveries are in rural areas where mothers have inadequate trained help, traditional practices are dominant (for example, putting various materials on the umbilical stump), and antenatal care is very limited. You would therefore:

- Train the traditional birth attendants.
- Educate them, and all the women of the community.
- Monitor and reform the antenatal clinics.
- Get a supply of tetanus toxoid.
- Ensure that all pregnant mothers at the clinics receive toxoid.

**Epidemic**

When an aetiological agent (toxic chemical, microbe) affects a population for a limited time, the incidence of new cases rises rapidly - this is described as an epidemic. It is, in fact, any marked increase in

**Attack Rates**

the incidence or prevalence of a disease. The attack rate can be measured in such a population for the period of the epidemic if the number of cases is related to the total population at risk. An attack rate is therefore a type of incidence rate (usually expressed as a percent), in which the time period is not mentioned. It is used for exposed populations observed for limited periods of time, such as during an epidemic.

$$\text{Attack rate} = \frac{\text{Number of cases occurring in a defined group}}{\text{Number of persons in the defined group}}$$

**Example**

For example, at a village feast 40 of 200 persons who ate a variety of foods became ill with diarrhoea within the next few days.

$$\text{Attack rate} = \frac{40}{200} \times 100 = 20 \text{ per } 100$$

Answer 1

Incidence rate $\frac{40}{500,000}$ = 0.00008

= 0.008 per 100
= 0.08 per 1000
= 8.0 per 100,000

**PREVALENCE**

Definition

Prevalence rates measure the number of people in a population who have a disease at a given point in time and is calculated thus:

$$\frac{\text{Total number of cases of a disease at a given time}}{\text{Total population}}$$

Period prevalence

If the time referred to in the numerator is a period of time, such as one year, then the rate is a period prevalence. An example of disease period prevalence would be the number of accident cases in the surgical ward of a hospital over a six-month period. Included in this is the number of new cases admitted during this period plus the old cases that were in the ward at the beginning of the six-month period.

Point prevalence

If the time referred to in the numerator is a specific time point, such as a particular date, then the rate is a point prevalence. An example of a disease point prevalence is the number of pulmonary tuberculosis cases in Kenya on 1st January 1987.

Uses

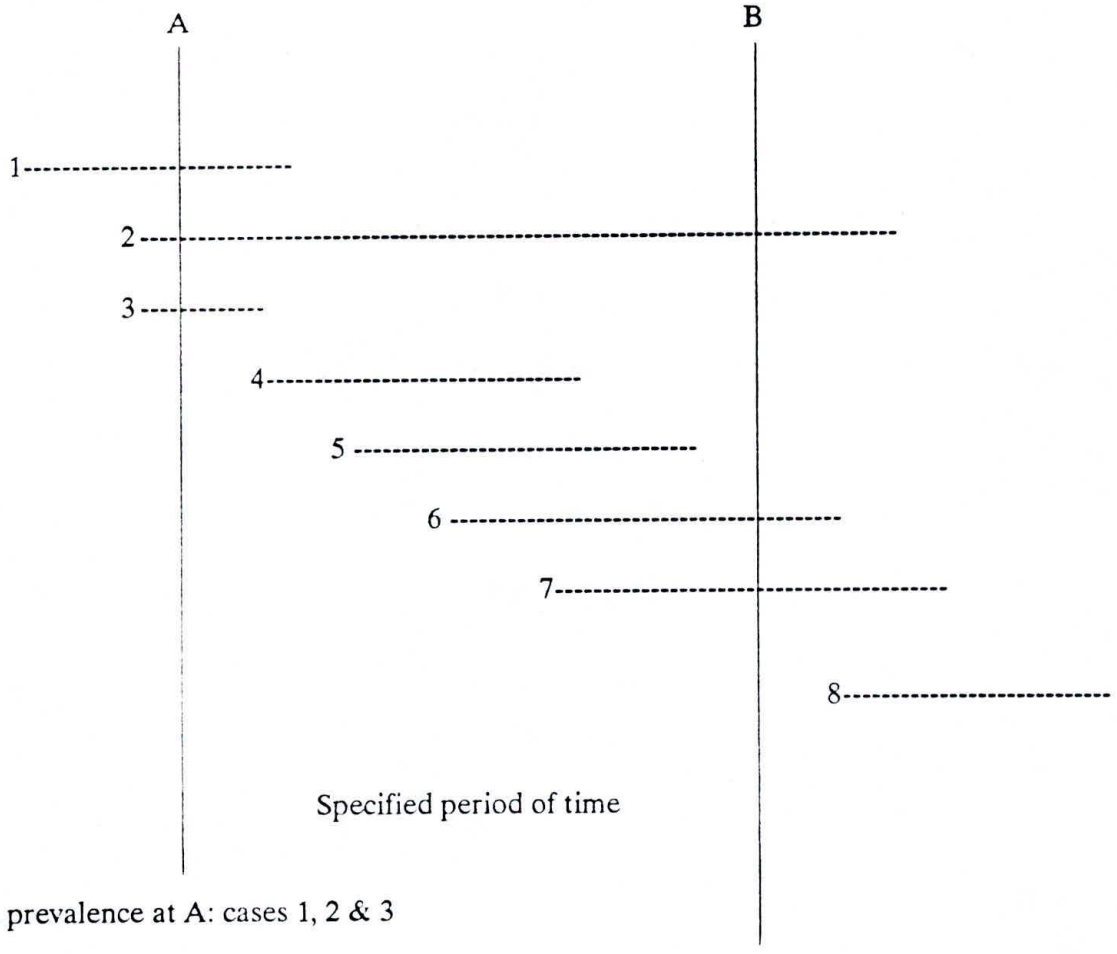
Prevalence is used by health planners because it measures the need for treatment and hospital beds, and aids in planning health facilities and manpower needs. It can be determined by a simple cross-sectional study (see Unit 3).

Incidence and prevalence

In your district there may be many cases of acute streptococcal pharyngitis - but each lasts only a few days. There is thus a high incidence but a low prevalence. By contrast the incidence of new cases of leprosy is low, but, because the disease continues for a long time (in contrast to streptococcal pharyngitis) the prevalence is high. Thus: prevalence = incidence x average duration.

A decrease in prevalence may result not only from a decrease in incidence but also from a shortening of the duration of disease, through either more rapid recovery (e.g. improvements in therapy) or more rapid death. Look now at Figure 1 which will help you to understand these terms without any doubt.

Figure 1 Incidence and prevalence




Point prevalence at A: cases 1, 2 & 3

Point prevalence at B: cases 2, 6, & 7

Incidence: cases 4, 5, 6 & 7

Period prevalence: cases 1, 2, 3, 4, 5, 6 & 7

**Denominators**

It is very important, when you are calculating rates, to choose the appropriate denominator. If you are estimating the prevalence rate of a particular disease, then you should use as the denominator the total number of individuals who may be at risk to this particular disease. In the case of a sample survey, this may comprise all individuals in the sample. However, if you wish to estimate the 'positivity rate' for *Plasmodium falciparum* from among the blood samples which have been collected in this survey, then you should use as the denominator the total number of individuals from whom blood samples were taken, not the total sample size. For age-specific rates, such as an age-specific mortality rate for children aged five to ten years, the denominator should include only children in the study sample who are aged between five and ten years.

**Samples**

**Specific rates**

Figure 4 Number of cases of pneumonia occurring in two populations, A and B

| | POPULATION A | | | POPULATION B | | |
|---|---|---|---|---|---|---|
| Age | Cases | Population | Incidence Rate/1000 | Cases | Population | Incidence Rate/1000 |
| 0-9 | 200 | 10,000 | 20 | 150 | 5,000 | 10 |
| 10-19 | 50 | 7,500 | 6.7 | 10 | 4,000 | 2.5 |
| 20-29 | 30 | 5,000 | 6 | 10 | 5,500 | 1.8 |
| 30-39 | 20 | 4,000 | 5 | 10 | 5,000 | 2 |
| 40-49 | 10 | 2,700 | 3.7 | 30 | 5,800 | 5.2 |
| 50-59 | 60 | 1,800 | 33.3 | 60 | 4,500 | 13.3 |
| 60 & over | 100 | 1,000 | 100 | 300 | 2,200 | 136.4 |
| Total | 470 | 32,000 | 14.7 | 470 | 32,000 | 14.7 |

*Exercise 2*

*Why do you need to calculate age-standardised incidence rates in order to compare the incidence of pneumonia in these two populations?*

.....

.....

.....

.....

.....

## STANDARDISATION OF INCIDENCE RATES

Comparing rates

In Unit 1 we discussed how mortality rates need to be standardised if we want to compare the rates in two different populations or the same population at different times. The same principles apply when we compare incidence rates.

Crude rates

Specific rates

When using rates to compare risks of disease it is important to consider whether the populations differ by a factor that is already known to affect the risk of the disease. You may already know that factors such as age, sex or race affect the rate of developing a particular disease. If the incidence of a disease is highest in very young children and low in adults, it would be inaccurate to compare the crude incidence rates of two populations which differed in their age composition. The calculation of age-specific rates would help you to make a proper comparison, but you may then have to compare a considerable number of figures.

Standardised rates

To solve this problem we can calculate a single summary rate figure, a standardised rate, that accounts or adjusts for the differences among populations regarding these other variables. That is, we adjust the rates to take into account the factors that are already known to greatly affect the risk of illness. Age is the most common factor that requires standardisation, but any factor known to have a large effect can be adjusted for.

Example

Let us look at an example[1]. Figure 2 shows the numbers of cases of a disease which occurred in two populations called A and B during a specified period of time.

Why standardise?

Firstly let us think about why standardisation is necessary. Both populations comprise 10,000 persons. The incidence of the disease in population B (18.2 per 1000) is much greater than in population A (12.5 per 1000). But, if you look at the table you can see firstly, that

Figure 2 Number of cases occurring in populations A and B

| | Population | Age group (years) 0-4 | 5-14 | 15+ | Total |
|---|---|---|---|---|---|
| A | No. cases | 63 | 50 | 12 | 125 |
| | No. population | 1500 | 2500 | 6000 | 10,000 |
| | Incidence/1000 | 42 | 20 | 2 | 12.5 |
| B | No. cases | 90 | 84 | 8 | 182 |
| | No. population | 2500 | 3500 | 4000 | 10,000 |
| | Incidence/1000 | 36 | 24 | 2 | 18.2 |

the incidence decreases considerably with age and secondly, that the age distribution in the two populations is very different, since B has a greater proportion of young people. Therefore, although B has more cases and a higher incidence than A, this may be the result of the

[1] From Barker, D.J.P., *Practical epidemiology*, Churchill Livingstone, 1982

different age structures of the two populations rather than a difference in the levels of exposure or susceptibility to the disease determinant.

Procedure for age standardisation

We can solve this problem through a procedure known as direct standardisation. The steps are as follows.

- Calculate the age-specific incidence rates. For example for 0-4 year olds in A, age-specific incidence = 63/1500 = 42 per 1000.
- Select a population with a 'standard' age distribution to replace A and B. Although any standard population can be used, its age distribution should approximate to that of A and B, and it is convenient to use one or other of them or their sum. In Figure 3 the two populations have been added to give a standard population (A+B) of 20,000.
- Multiply the age-specific incidences in A and B by the standardised numbers of cases. For example, among 0-4 year olds in A, the standardised number of cases is (42 x 4000)/1000 = 168 (Figure 3).

Figure 3 Numbers of cases resulting from application of incidences in A and B to standard population

| | Age group (years) | | | Total |
|---|---|---|---|---|
| | 0-4 | 5-14 | 15+ | |
| No. standard population (A+B) | 4000 | 6000 | 10,000 | 20,000 |
| No. cases in A | 168 | 120 | 20 | 308 |
| No. cases in B | 144 | 144 | 20 | 308 |

The total of standardised numbers of cases in both A and B is 308, giving an age-standardised incidence of 308/20,000 = 15.4 per 1000 for each population.

Therefore, the difference between the unstandardised or crude incidence rates of 12.5 and 18.2 per 1000 in A and B is solely attributable to their different age structures.

Age/sex standardised rates

In this case we standardised for age only. If you need to standardise for sex as well as age the procedure is identical to the one we have just described, with the addition that numbers of cases in the standard population are calculated for males and females separately, and are then combined to give an age/sex-standardised rate.

Now try the following exercises.

*Exercise 3*

*Standardise the incidence rates using the method of direct standardisation. What do you conclude about the incidence of pneumonia in these two populations?*

Answer 2

Both populations are the same size (32,000) and both have the same incidence rate (14.7). However, the population structure of A and B are very different: A has more young children and less old people than B. If you want to make an accurate comparison of the incidence rates you must standardise the two populations by calculating age-standardised incidence rates.

Answer 3

Figure 5 summarises the steps you should have followed to calculate the age-standardised incidence rates.

Figure 5 Steps in direct standardisation

| | Population A | | | Population B | |
|---|---|---|---|---|---|
| Age groups | Age-specific incidence | Stand No. cases | Stand Pop. A + B | Age-specific incidence | Stand No. cases |
| 0-9 | 20 | 300 | 15,000 | 10 | 150 |
| 10-19 | 6.7 | 77 | 11,500 | 2.5 | 29 |
| 20-29 | 6.0 | 63 | 10,500 | 1.8 | 19 |
| 30-39 | 5.0 | 45 | 9,000 | 2.0 | 18 |
| 40-49 | 3.7 | 31 | 8,000 | 5.2 | 44 |
| 50-59 | 33.3 | 210 | 6,300 | 13.3 | 84 |
| 60+ | 100 | 320 | 3,200 | 136.4 | 436 |
| Total | 16.3 | 1,046 | 64,000 | 12.2 | 780 |

Thus the age-standardised incidence rate is 16.3/1000 for population A and 12.2/1000 for population B. You can now see clearly that population A has a higher incidence of pneumonia than population B, and that the similarity in their crude incidence rates was entirely due to the different age structures of the two populations.

## 2.2 MEASURES OF RISK

In the community some people will smoke, others will not; some mothers will attend for antenatal care, others will avoid it; some may have piped water while others do not. It is therefore important for accurate health education, planning and care, that risks should be known. Such risks can be derived if the incidence rates (see section 2.1) in an exposed and an unexposed population are known.

### RELATIVE AND ATTRIBUTABLE RISK

Absolute risk

These are calculated from the absolute risk which is the rate of occurrence, or the incidence of a condition. They are two measures of the association between exposure to a particular factor and the risk of a certain outcome (e.g. disease).

**Relative risk**

Definition

In our daily practice we use relative risk, which is the ratio of the incidence rate among the exposed to the incidence rate among those not exposed to the factor. It is not itself a rate, and therefore does not indicate the incidence of disease, but we can learn from it how much the risk associated with a certain factor (for example, smoking) is increased for an individual patient.

$$\text{Relative risk} = \frac{\text{Incidence rate among the exposed}}{\text{Incidence rate among those not exposed}}$$

Uses

Clearly if the relative risk of a factor is high among a group of people, they will benefit greatly if that factor is removed. But the relative risk gives no indication whatever of the overall prevalence of a factor in the population and so is not a guide for major preventive programmes. A factor needs to have both a high relative risk and be prevalent in the population in order for it to influence the disease in the population. A high relative risk suggests that a factor is significant in the cause of a disease and so it may be a useful pointer towards research to define how it operates. (Refer to Units 3 and 6 for information on analytical studies designed to investigate cause).

Interpretation

**Attributable risk**

Definition

This measures the amount of the absolute risk (incidence) which can be attributed to one particular factor (e.g. smoking). It is measured by simple subtraction of the incidence rate among those not exposed to a factor (non-smokers) from the rate among those exposed to a factor (smokers), i.e. the attributable risk is the excess incidence among the exposed.

$$\text{Attributable risk} = \text{Incidence rate among exposed} - \text{Incidence rate among non-exposed}$$

As defined above, the attributable risk indicates the excess of disease due to a factor in that subgroup of the population which is exposed to the factor. If we replace 'incidence rate among exposed' in the formula for attributable risk with 'incidence rate in the total population', we have the population attributable risk. The population attributable risk is generally of significance to public health planners as it measures the potential benefit to be expected if the exposure could be reduced in the population. If you can identify a high population attributable risk you have an opportunity for removing a significant factor in disease from your community.

Example

Let us take the example of low birthweight and lack of antenatal care: if you find a high attributable risk you can get your team to work with the community leaders so that antenatal care becomes routine and normal, not irregular and artificial.

Uses

A high attributable risk relates to a factor in the population at large and so is valuable in health care planning and education. A high relative risk relates a factor to a particular group in the population, and so is valuable for the individual patient and as a basis for a research hypothesis.

One final note: risk estimates are probability statements and it must be remembered that:

- All those exposed to the factor do not develop the disease, they merely have an increased probability of doing so.
- Some people who have not been exposed to the factor will develop the disease.

Refer to Unit 6 for more information on association and causation.

---

## 2.3 EPIDEMICS: AN ILLUSTRATIVE EXAMPLE

### INVESTIGATION OF AN OUTBREAK

We shall study the theory of epidemics from this example as it is just the sort of problem that you could face in your district. Read the following case study.

Case study

A *A medical officer who has charge of a district in which there is a village of about 2000 people is told that there is an outbreak of diarrhoea there. Two old women have died, the health clinic/aid post can not cope with all those who are affected, and he is therefore asked to do something quickly. [He has inadequate laboratory facilities in his health centre for culture of stools].*

B *He takes action and the number of cases falls off rapidly but then more cases are seen, one or two at a time over a period of three weeks, gradually increasing to a maximum of 30 cases in one week. Then, quite quickly there are very few cases reported. There is a good health assistant in the clinic aid post and he has got good data when the medical officer visits him.*

Let us first consider what happened in the outbreak described in paragraph A.

*Question*

*What essential data would you ask for if you went to the village? Write down a list of questions you would ask the health assistant. Then and look at our suggestions on the next page.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

I would ask three questions.

1 How many cases were reported each day?
2 What is the age and sex distribution of the cases?
3 Has the health assistant been able to identify the homes of those affected?

Data

In answer to your first question the health assistant gives you the following data about the number of cases.

| Day | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| No. of cases | 3 | 13 | 12 | 2 | 1 | 0 |

Notice the rapid rise in the number of cases and the rapid fall also.

Point-source outbreak

This is typical of one type of outbreak or epidemic: a point-source epidemic/outbreak. This is where a simultaneous exposure of many susceptibles to a source of a pathogenic agent results in an explosive increase in the number of cases of the disease over a short time.

Explanation

This type of pattern is typical of water- and food-borne diseases (for example cholera, typhoid, food poisoning outbreaks) and those where the source is a common fomite.

The picture may be modified in outbreaks where the source provides continuous exposure over a period of time (extended point-source outbreak). The onset is still abrupt but the cases are spread over a greater period of time.

The point-source outbreak and its extended form are illustrated in Figures 6 and 7.

Figure 6 Point-source outbreak

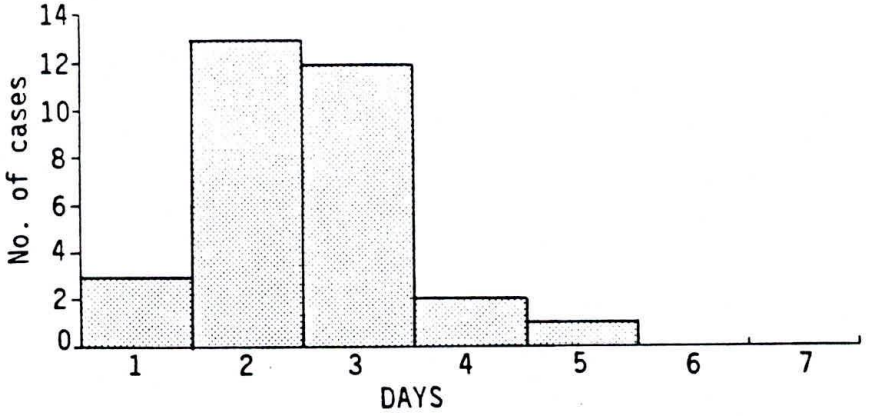
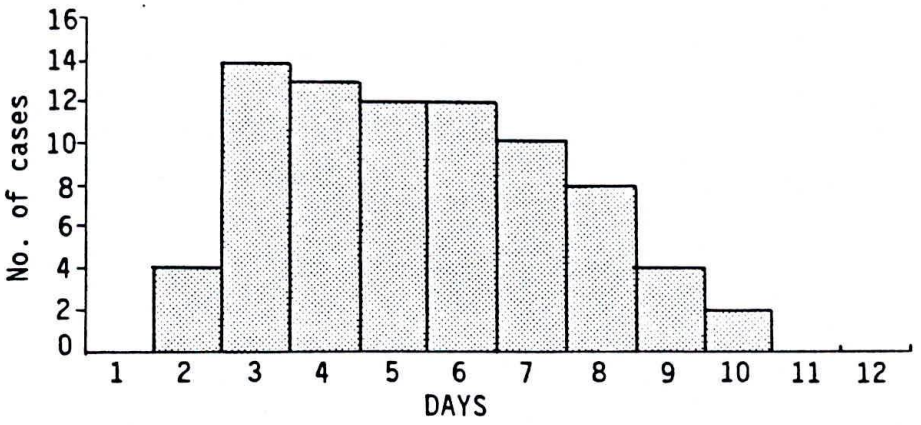



Figure 7 Extended point-source outbreak



Age and sex distribution

In answer to your second question (age and sex distribution of cases), the health assistant tells you that the first five cases were all adults but, after that men, women and children were equally affected. But there was no case in breast-fed infants.

*Question*

*What can you conclude from this?*

.....

.....

.....

.....

Transmission

- That the causative pathogen was spread among adults from a source common to them and that it thereafter spread to other members of their families.
- That the pathogen was in something which breast-fed infants did not receive.

This is just what the health assistant found and, as he had been well taught and was an enterprising man, he quickly identified the homes of those affected (your third question).

Affected homes

In brief, the first five cases were adults from the same compound but then three children and three more adults were affected from the very same compound. An adjacent compound had one case and the others were spread randomly through the village.

*Question*

*What would you do with these data?*

.....

.....

.....

.....

Action

These are our suggestions. First I would construct a map. Then I would examine the map and look for evidence of clustering in the village (or even in the compounds) because this could lead to my identifying the source of the outbreak. If I could identify such a focus, I would go to the affected homes/compound(s) where I would interview all those affected: this would probably enable me to limit the source to an individual or a common food or water source.

Source of infection

To your dismay you find an unprotected well, and just above it in the compound is the pit latrine which is poorly made. It is the rainy season and you conclude that the family's water has been contaminated by faeces from the poor pit latrine, 'waterwashed' by the heavy rains.

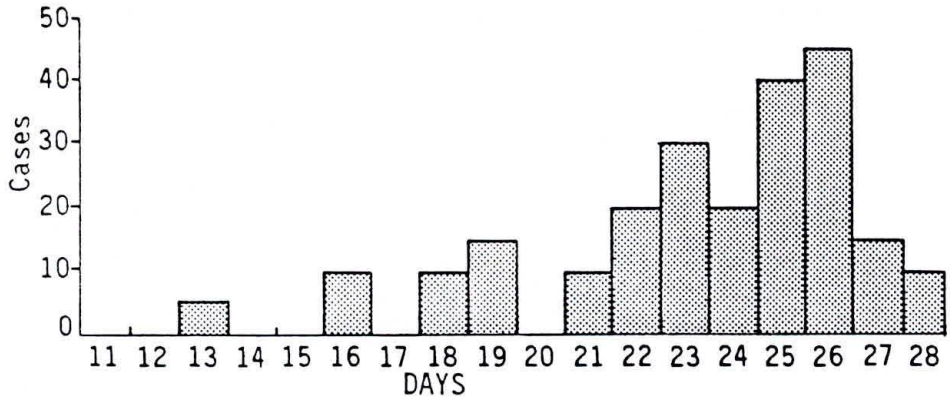
Cholera

The health assistant has already concluded that there must be a source of infection in this compound which he had therefore visited. Because the cases were typical of cholera he had treated the well with chloride of lime to kill *Vibrio cholerae*. He had already had experience of cholera in his previous posting: he was familiar with the phenomenon of 'clustering' of cases; with the need to define a source if possible; and with basic prevention by putting a bactericidal agent in the well.

Propagated-source outbreak

Let us now consider what happened in the outbreak described in paragraph B (page 59). In this section, what appeared to be a point -source outbreak has turned into a propagated-source outbreak. This is an outbreak in which the organism is propagated in the community by passage from person to person, so that the initial rise in the number of cases is less abrupt than in point-source outbreaks, and the decline is usually less gradual (Figure 8).

Figure 8 Propagated-source outbreak




From Figure 8, it is clear that transmission is continuing after the initial explosive outbreak.

**FACTORS DETERMINING THE PATTERN OF AN OUTBREAK**

So much, then, for our definitions and description of types of outbreak/epidemic. Let us now return to the epidemic of cholera which we are studying. In this outbreak there was a short incubation period and common contact with an infective source for all the victims. This is a useful model for any acute outbreak of disease, and we can use it to think of the factors which determine the pattern of any outbreak.

*Question*

*What factors do you think could determine the pattern of an outbreak? Think in terms of the microbe, the people and the transmission of the microbe to the people.*

.....

.....

The following factors may determine the pattern of an outbreak:

Microbe
- Pathogenic microbe.
- Incubation period - if short, explosive outbreak.

Vector
- If vector-borne - favourable conditions for vector
  - duration of maturation of microbe in vector.

People
- Population - susceptible.

Microbe/people
- Transmission easy - overcrowding (contact)
  - common food/water (faeco-oral).

**Pathogenic microbe**

We shall now use a graph/histogram of the time/incidence of the infection to analyse the behaviour of the microbe and how it can help us to define the organism and its victims more precisely. Look back at Figure 6: we can use it in two ways.

**Time of exposure unknown - organism known**

Defining the source

If we use *V. cholerae* as an example, we could redraw the figure to show an incubation period of two to three days before day two on the histogram (which shows the onset of clinical symptoms). Therefore, if the movement of people is known we can find out where the victims were at the time they were infected. This is a much more accurate way of analysing a point-source outbreak. A map of the place where each was infected can be made.

**Time of exposure known - organism unknown**

Defining the organism

For example, if you are investigating a food-poisoning outbreak and the time the meal had been taken was known, then the incubation period can be calculated. This will give a clue to the infecting organism:

- 4-8 hours *Staphylococcus*
- 24 hours *Salmonella*
- 48+ hours *V. cholerae*

**Population characteristics**

If we look at the characteristics of people who are affected by the epidemic, such as age, sex or occupation, this may give us some clues to the source of the outbreak. For example if the initial cases are in schoolchildren, look for a source of infection at school.

**Population size**

Immunity

Herd immunity

As has been discussed above, the determinant in the continuation of an epidemic is the number of susceptible people who remain in the population. Once an individual has experienced an episode of the disease (whether manifest or not) he may develop immunity (either temporary or permanent). When a certain number of individuals have developed immunity then there are insufficient susceptibles and the disease dies out. This immunity (if it is a permanent immunity as occurs in viral infections) is carried on as the herd immunity. After a period of time, depending on the size of the population, this herd immunity becomes diluted sufficiently by new individuals born (or by

immigration) into the community, and a new epidemic can take place.

Critical population

The population size that can overcome the epidemic threshold is called the critical population. This is the theoretical minimum host population size required to maintain an infecting agent. It depends upon the infectious agent, the demographic structure and the conditions (hygiene, etc.) of the host population. In developing countries with immense birth rates the critical population becomes less and the frequency of epidemics increases. Examples of the critical human population size are 500,000 for measles and 10,000 for varicella.

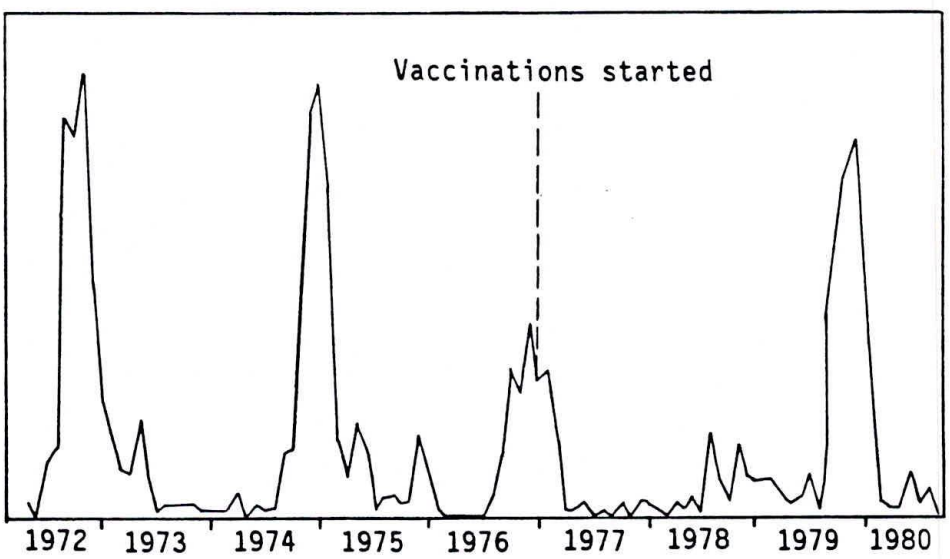
Regular epidemics

If the population is less than the critical size, then regular epidemics will occur at intervals dependent on the population size. An example is given in Figure 9 of a measles epidemic which occurred regularly every two years in a well defined community. These regular epidemics can be analysed in the same way as a propagated source epidemic, from which it has been shown that the smaller the community, the longer is the interval between epidemics.

Critical rate of vaccination coverage

An extension of the concept of herd immunity shows that not everyone in a population needs to be vaccinated to prevent an epidemic. On the same principle as calculating the critical population, the critical rate of vaccination coverage can also be worked out. It can similarly be shown that even if this target is not reached, then the epidemic will be put off until a future date when the susceptible unvaccinated children will have grown older and therefore be able to cope with the infection better. This is also illustrated in Figure 9.

Figure 9 Prolongation of the time interval between measles epidemics due to vaccination (Namanyere, Tanzania)



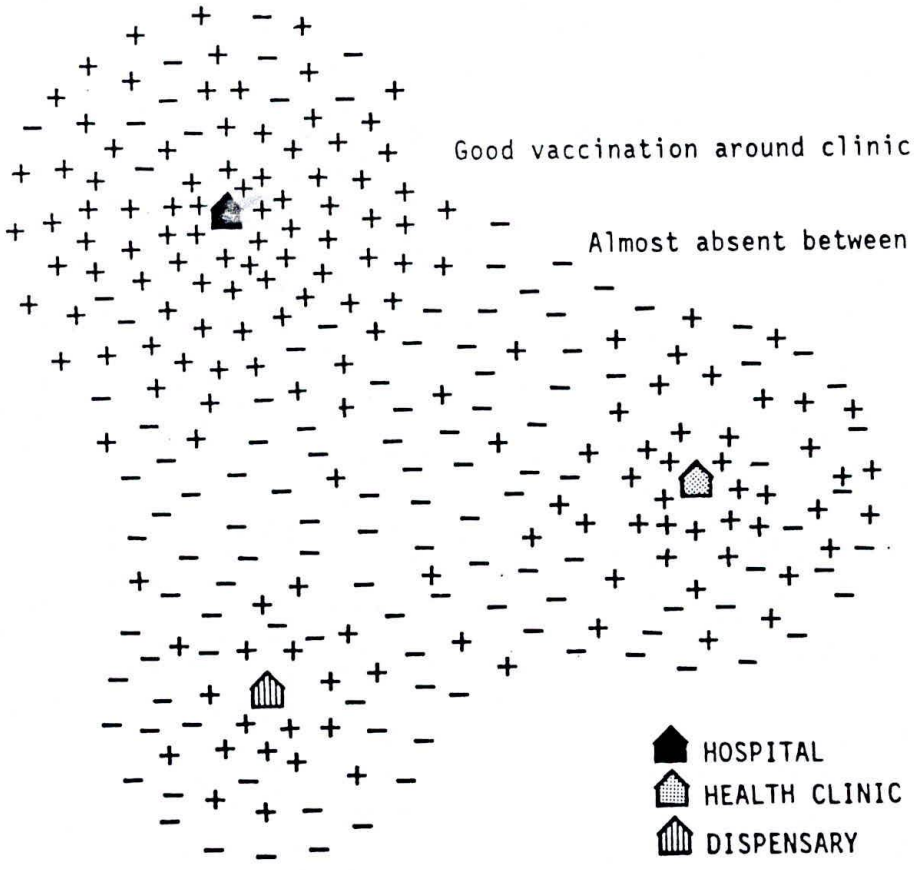
Unequal vaccination coverage

Unfortunately, this theoretical argument depends on uniform vaccination coverage (even of the lower rate), which does not occur. In the smallpox vaccination programme it was calculated that 80% vaccination coverage would be sufficient to prevent transmission, but when this was achieved, the infection still continued. What had happened was that, while the overall average was 80%, in some remote areas it was as low as 50 or 60%, but this was matched by large urban areas where 90% cover was achieved. The 80% average therefore concealed the true position in the rural areas so that smallpox transmission inevitably continued in these remote areas.

Effective vaccination coverage

Vaccinations can either be done by mobile teams, or static health units. To be effective both systems are required and have to work in coordination with each other. If static units only are considered (which is a common pattern) it will be noted that the vaccination coverage radiates out from the unit in a target fashion. This can be schematically illustrated for three health units. (Figure 10).

Figure 10 Unequal vaccination coverage from static clinics




Around each health unit there is a good coverage of vaccination, but between them it is almost nil. This can be overcome either by making the health unit cover a wider area, or by using mobile clinics to fill in the gaps in vaccination programmes.

## FURTHER ANALYSIS OF AN OUTBREAK

If you wish to confirm the source of an infection which has been suggested by your descriptive investigation, you may need to do a further analytical study (see Unit 3). There are two possible approaches: a case-control study or a cohort study.

Case-control study

Let us think how we can investigate three possible sources of infected water in the outbreak of cholera using a case-control study. We question all those affected about the food/water taken during the previous two to three days and then try to identify which of the sources harbours the pathogen. In order to make sense of the information we must ask the same questions of a group of controls (people from the same area who do not have the infection). You might then get the information shown in Figure 11.

[1]Figure 11 Case-control study of a cholera outbreak

| | Total | Source of water | | |
|---|---|---|---|---|
| | | Well 1 | Well 2 | Stream |
| Cases of cholera | 18 | 17 | 16 | 6 |
| Controls healthy | 18 | 14 | 3 | 17 |

*Question*

*Can you identify which source of water is infected - well 1, 2 or stream?*

.....

.....

If only the 18 cases had been questioned (i.e. no controls used), the likely source of infection could have been either well 1 or well 2. However, far fewer of the controls had used well 2, so it is reasonable to assume that this was the source of infection.

Cohort study

An alternative analytical approach is to use the cohort form of enquiry in which persons exposed and not exposed to the possible source are compared for frequency of the disease.

This technique is particularly useful in a food-poisoning outbreak, in which the meal responsible is known but not the item of food responsible. The different attack rates of illness are calculated and compared for those who ate each item of food and for those who did not eat the item. This is illustrated in Figure 12 where it can be seen that rice is the likely culprit.

[1] Adapted from: *Epidemiology for the health officer: a field manual for the tropics.* Edited by W.O. Phoon. WHO, 1985

[1]Figure 12 Food-specific attack rates of food poisoning outbreak

| | No. ate specified food | | | No. did not eat specified food | | |
|---|---|---|---|---|---|---|
| | Ill | Not Ill | % Ill | Ill | Not Ill | % Ill |
| Lunch (food eaten by 161 persons) | | | | | | |
| Chicken | 111 | 39 | 74.0 | 5 | 6 | 45.5 |
| Beans | 106 | 41 | 72.1 | 10 | 4 | 71.4 |
| Potatoes | 80 | 30 | 72.7 | 36 | 15 | 70.6 |
| Rice | 116 | 40 | 74.4 | 1 | 4 | 20.0 |
| Pineapple | 109 | 39 | 73.6 | 7 | 6 | 53.9 |

Significance tests may be used to test your hypothesis (see Unit 7). The one ill person who did not give a history of eating rice may well have forgotten that he did take rice, or alternatively, some cross contamination may have occurred. Again it is the difference between attack rates (percent ill) that is important.

## CONTROL OF AN OUTBREAK/EPIDEMIC

All this theory about epidemics, which we have tried to make as relevant as possible, is designed so that you can act more effectively in an epidemic or an outbreak.

Whatever the outbreak, the following steps should be taken:

Principles of control

- Identify and record all the cases.
- Trace the source of the infection.
- Attack the source, and block transmission.
- Treat, and 'neutralise' from infecting others, all cases.
- Immunise population (if applicable).
- Practise surveillance.
- Work through the community leaders to educate the people about the cause of the outbreak.

[1] Adapted from: *Epidemiology for the health officer: a field manual for the tropics.* Edited by W.O. Phoon. WHO, 1985

# UNIT 3: OBJECTIVES

Study with this unit will enable you to plan a simple epidemiological investigation in your district.

In particular you will be able to:

- Formulate objectives and hypotheses appropriate to the purpose of the investigation.
- Select a study design appropriate to the investigation.
- Choose a study population and a method of sampling it.
- Assess the need for control groups and methods for their selection.
- Define the observations to be made and choose a standardised technique for making these observations.
- Identify and minimise potential sources of bias, error and variation.
- Determine what background information, resources and administrative procedures are needed in order to carry out the investigation.

---

# UNIT 3: CONTENTS

## 3.1 INTRODUCTION

We can recognise three main stages in the planning of an epidemiological study:

- Deciding whether an investigation is needed.
- Deciding what methods are to be used.
- Deciding what resources are needed.

Why do a study?

You will first want to know why you might need to do an epidemiological study. In fact, you may encounter many situations in which you may see a need for a study, but the underlying reason is so that you can acquire essential health information and make the best use of it in the conduct of the health care activities under your responsibility. Epidemiology is about information: the information needed for health planning, and the supervision and evaluation of health promotion and disease control activities. You may be able to obtain the information you need from existing reports and records (see Unit 1), and these sources of information may help you to identify particular health problems. However, there will be occasions when you need to collect new data to answer specific questions, or to help you plan your health services more efficiently.

Information

Uses of studies

We can recognise four broad categories of studies:

- Periodic or continuing assessment of the health status of the community (a health survey).
- Periodic or continuing evaluation of the effectiveness of health services.
- Surveillance for specific diseases.
- Investigation and containment of disease outbreaks.

These last two, surveillance and epidemics, were discussed Unit 2. Let us think a little about the uses of a health survey. These can be broadly grouped as follows:

Uses of health surveys

- To measure the total amount of illness in the population.
- To measure the amount of illness caused by a specified disease.
- To study the nutritional status of the population.
- To examine the utilisation of existing health care facilities and the demand for new ones.
- To measure the distribution in the population of a particular characteristic, such as breast-feeding practice, haemoglobin levels, sources of drinking water, etc.
- To examine the role and relationship of one or more factors in the aetiology of a disease.

A study has to be well-planned if the results from it are to be accurate and rewarding. Hasty and slipshod planning is likely to produce worthless results. This unit will help you through all the stages in planning so that you can be sure that your findings will be of maximum use to yourself and to others.

Stages of a study

Let us begin by getting an overall picture of what is involved in doing an epidemiological study. Look at Figure 1 which shows the various stages.

In this unit we will be considering all those steps which lead up to the implementation of a study. Subsequent units will deal with data collection (Unit 4), processing and analysing data (Units 5, 6 and 7), and presenting data and writing a report (Unit 8).

This unit will help you to work systematically through the stages involved in planning an epidemiological study, although you should realise that the various elements in planning are interdependent and are not completely separate steps. We will pay a great deal of

Sound planning

attention to the various aspects of sound planning, such as the careful choice of the study population and of definitions, and the use of standardised and accurate methods of collecting information. The more carefully your investigation is planned the more likely you are to obtain useful findings.

Imperfections

However, it is very difficult to produce an absolutely perfect study. Almost invariably practical difficulties, oversights and accidents produce imperfections in the methodology. What is important is that you should be aware of these imperfections, examine their impact and take them into account when interpreting your findings. If this is done, your study should be a sound and useful one.

Literature

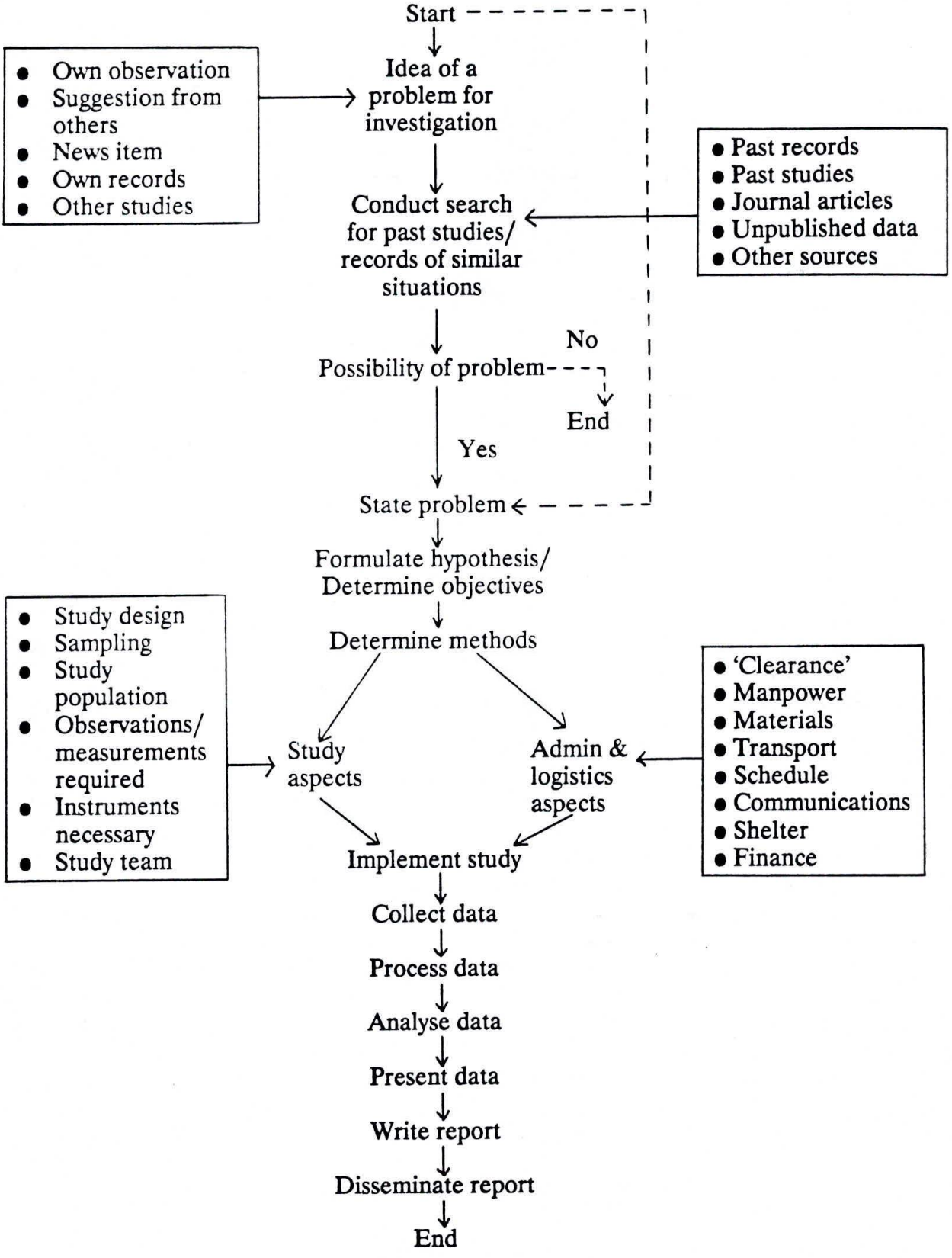
One final word, 'reviewing the literature' is not discussed specifically in this unit, not because it is unimportant, but because it should be done throughout the study if at all possible. Published data, as well as the experiences and thoughts of others may indicate the presence and the nature of the research problem. They may also be of great help in every aspect of planning and in the interpretation of the findings.

Discussions

Even if it is difficult for you to obtain published literature, discuss your investigation with your colleagues and your supervisor, if you have one.

---

Figure 1 Conduct of an epidemiological investigation




Adapted from: *Epidemiology for the health officer: a field manual for the tropics*. Edited by W.O. Phoon. WHO.

## 3.2 WHAT AM I TRYING TO FIND OUT?

### KEY POINTS

The main issues involved in this stage of planning an epidemiological study are summarised below and then discussed in more detail in this section.

Purpose
Objectives
- Define the purpose of the study (what is the study problem?).
- State the objectives of the study (what are the questions to be answered?).

Hypothesis
- If you are testing a theory (for example about the aetiology of a disease or cause of an outbreak), you will need to formulate specific hypotheses.

Analysis
- Decide in general terms how you will analyse the results.

Data needed
- Decide what information you have already and what you need to collect in order to achieve the objectives or test the hypotheses.

Constraints
- Decide what you can in fact do, given the constraints of your resources.

### PURPOSE

Clarify the purpose

The first step in planning a study is to clarify its purpose. For example, is the purpose to obtain information which will help you to decide how to use your resources? Or is the purpose to identify persons who are at special risk of contracting a specific disease, so that you can take preventive action? Or is the purpose to study a specific aspect of aetiology?

Importance

Whatever the purpose, the reasons for doing a study must be clearly defined in your mind, and also on paper as the first part of your study plan or protocol. It is impossible to plan any study unless you are clear about its purpose. An epidemiological investigation requires a considerable amount of time and resources, so you must ensure that the information you are looking for will really be useful, and that the information you obtain will be used.

### FORMULATING OBJECTIVES

When you have decided what to study and why you want to study it, you can formulate your study objectives. That is, you can state what knowledge you want the study to provide - what questions you are trying to answer.

Phrasing objectives

Objectives can be phrased quite simply as statements or questions. For example:

'What is the infant mortality rate in population Y during period Z?'
'What is the incidence of burns in children?'
'What is the case fatality rate of measles?'.

Even in quite simple studies it may be important to obtain separate information for different groups in the population, perhaps for specific age groups, for the two sexes or for different cultural groups. In such studies, the objective might be stated as:
'To measure X in population Y by age, sex, etc.'

Specific and simple

The more specific and simple the objectives of a study, the more definite will be the results. Avoid the temptation of collecting additional information which is not related to the study objectives. If your purpose is to study vaccination status, do not try to record malnutrition as well. If you try to do too much, nothing will be done well and the entire survey will be of little use.

How achievable?

When you have stated your objectives, examine them carefully to make sure they are achievable given the resources of the study (see below) and availability of data. Check first whether information on some of the objectives is already available.

Associations

Surveys which collect information on the health of a community may reveal associations which provide pointers to the different health needs of different groups in the population. They may show, for example, that a disease is more common in people who eat a particular diet. You may then wish to do a specific study to find out more about this relationship. Or you may want to examine the role of one or more factors in the aetiology of a disease. If you want to investigate the relationships between various factors, or to look for cause, you will need to ask more specific questions. The way this is usually done is to make statements (hypotheses) about what we believe the relationships to be, and then to test these hypotheses.

Aetiology

## FORMULATING HYPOTHESES

An hypothesis

An hypothesis is a guess (or conjecture) about the nature of some process or interaction which is tested by collecting facts which leads to its acceptance or rejection. Scientific thought depends on the formation of hypotheses. Instead of collecting data with the aim of 'let's see what we get', we should collect data with the aim of disproving an hypothesis.

A research hypothesis may be stated as a positive declaration:
'The infant mortality rates in regions A and B are different.'
or 'The rate is higher in region A than in region B.'

Null hypotheses

or it may be stated as a negative declaration (a null hypothesis):
'There is no difference between the rates...'
or 'The rate is not higher in region A than in region B.'

Testing an hypothesis

The essential principle of any well formulated hypothesis is that it is refutable. That is, there exists some practical means of demonstrating that it is incorrect. All hypotheses must be formulated with this quality firmly in mind. You might think that it would be better to **confirm** an hypothesis rather than disprove it, but in fact it is extremely difficult to prove that an hypothesis is right. It would, for example, be technically much easier to prove that 'there are **no** differences between males and females' (the null hypothesis) than it would be to prove that 'males and females **are** different'. Another reason for using a null hypothesis is that if you try to confirm an

hypothesis it is very tempting to look for evidence which supports your hypothesis, and to overlook contradictory evidence.

Analysis

Hypotheses are needed if you are going to analyse your results statistically, and it is important at this stage to think ahead to how you will need to analyse the results in order to test the hypotheses you have formulated. The analysis of studies designed to investigate associations and cause is considered in Units 6 and 7. Ask for help from your supervisor or an experienced epidemiologist or statistician if you have problems.

Thinking about the analysis is an important part of planning an epidemiological study. Do not wait until you have collected all the data before deciding how you are going to analyse the results in order to answer your study questions or test your hypotheses.

## PRACTICAL CONSIDERATIONS

When you are formulating your study objectives make sure that what you propose to do will be possible in practice.

Constraints

What you can do will be limited by the available staff time and skills, and material resources as well as your own ability, expertise and time. You should make an initial assessment of how much money, manpower, time and other resources are available before you begin serious planning. We shall consider resource requirements in more detail in section 3.6, but you should remember the limits of your resources throughout the planning stage.

A modest, well-planned study will be of more value to you, your colleagues and the community, than will an ambitious project which places too great a demand on you and your resources.

Writing objectives

We strongly recommend that you formulate your study objectives and hypotheses in writing. They should satisfy three requirements:

- Do they meet the purpose of the study?
- Are they phrased clearly and unambiguously?
- Are they expressed in measurable terms?

That is, the objectives should be realistic (answerable questions, testable hypotheses) and formulated in operational terms which can be applied in practice.

---

## 3.3 WHAT TYPE OF STUDY SHOULD I CHOOSE?

### KEY POINTS

The main points to be considered at this stage of planning an epidemiological study are summarized below. These will then be discussed in more detail in this section.

Choice
- The choice of study type (study design) depends on the purpose, objectives and hypotheses of the study.

Study design
- There are two main types of study design:
  - Descriptive studies are designed to collect facts.
  - Analytical studies are designed to test hypotheses and examine cause and effect relationships.

  Many studies have elements of both types.

Descriptive studies
- Descriptive studies may be further subdivided into cross-sectional (measuring prevalence) and longitudinal (incidence).

Analytical studies
- Analytical studies are of three kinds:
  - Case-control or retrospective;
  - Cohort or prospective;
  - Intervention studies.

Controls
- Analytical studies require the use of thoughtfully chosen comparison groups or controls. They must be carefully designed in order to reach clear and correct conclusions about the cause-and-effect relationships which are studied.

### STUDY DESIGNS

The objectives and hypotheses which you formulated (section 3.2) will determine the study design. In this section we shall examine the various types of study design so that you will be able to choose one which is appropriate for the problem you wish to investigate.

There are two main types of study: descriptive and analytical.

Descriptive studies

Descriptive studies are designed to **describe** certain characteristics of a problem (time, persons, place), by collecting information on specific problems or diseases in specified population groups.

Descriptive studies may be:

Cross-sectional
- Cross-sectional, which investigate individuals at a single point in time, thus giving estimates of point prevalence (see Unit 2).

Longitudinal
- Longitudinal, which measure individuals over a period of time, thus giving estimates of incidence (see Unit 2). Longitudinal studies may either be prospective (looking forward), measuring individuals as they become ill or recover, or retrospective (looking back), studying events of illness that occurred in the study population in the past.

The information which is obtained may enable us to develop hypotheses about the pattern of disease. These can then be tested using an analytical study design.

Analytical studies

Analytical studies are designed to **explain** the distribution of disease in specified population groups by testing hypotheses. These hypotheses may be derived from a descriptive study (done either by you or someone else), clinical observations or an examination of existing records. For example, early studies of the problem of the association between cigarette smoking and lung cancer were essentially descriptive, while a comparison of smoking rates in lung cancer patients and in equivalent patients without lung cancer is analytical. Analytical studies attempt to show whether particular events (such as the consumption of certain foods) or states (such as malnutrition or living in overcrowded conditions) act as 'causes' whose 'effect' is the resulting disease. Analytical studies are therefore based on the comparison of two or more groups of individuals. Three kinds of comparisons can be made and there are correspondingly three types of analytical studies.

Case-control studies

- Case-control or retrospective studies compare the characteristics of persons with disease (cases) with persons without the disease (controls), to find out whether the suspected determinant (cause) occurs more frequently among persons with the disease. Cases and controls must be comparable in all aspects except exposure to the suspected determinant of disease.

Cohort studies

- Cohort or prospective studies compare over a period of time the characteristics of persons exposed to the determinant (study population) and those not exposed to the determinant (controls) to find out whether a greater proportion of those exposed to the determinant develop the disease. (A cohort is a group of individuals who share a common experience or attribute).

Intervention studies

- Intervention studies aim to study the effects of active intervention which results in either the exposure or the prevention of exposure of a group of people to some determinant or risk factor. The effects of the intervention on the study group are determined by comparing this group with a similar group of controls who did not receive the intervention. The evaluation of a disease control programme is a type of intervention study.

Distinguishing case-control and cohort studies

Some people find it difficult to understand the difference between case-control and cohort studies. The following example may help to clarify the difference.

To test the hypothesis that smoking is associated with lung cancer using a case-control study you would begin with a group of patients with lung cancer (cases) and a group of patients without lung cancer (controls). You would then retrospectively determine the numbers of smokers and non-smokers in both case and control groups.

To test your hypothesis that smoking is associated with lung cancer using a cohort study, you would begin with a cohort of persons who are either smokers or non-smokers. You would then follow this cohort prospectively to determine how many smokers and non-smokers eventually developed lung cancer.

Analysis

If you are still unsure about the difference between these two types of study, refer to Unit 6 which describes how the results of

these studies are analysed. In any case, if you are planning an analytical study you must understand how the results should be analysed and what information the analysis will be able to give you.

Summary

Your choice of study design will depend on the following factors.

- The purpose of the study, its objectives and hypotheses.
- Manpower, financial and time constraints.
- The amount of expertise available.
- The quality of existing records.

Examples

Below are examples of epidemiological investigations. When you have read them, try to answer the exercise questions which follow.

A Between June 1974 and June 1977, visits were made each fortnight to 4000 households to find out whether any of the children had been ill with diarrhoea since the previous visit.

B A study was carried out on three groups of village children which were followed up from birth to three years of age. Two of the groups received weekly antimalarial chemoprophylaxis throughout the observation period, while children in the third group were treated only when 'clinical illness' occurred.

C Stool samples collected from 280 schoolchildren in five villages were examined for *Schistosoma mansoni* in August 1982.

D During an investigation into an outbreak of cholera, people who had developed the disease and a similar group who had remained healthy were interviewed as to whether they had eaten raw fish or not.

*Exercise 1*

*For each of the examples (A, B, C, D) identify the purpose of the study and the study design.*

A.....

.....

.....

.....

B.....

.....

.....

C.....

.....

.....

3.3 What type of study....?

D.....

.....

.....

*Exercise 2*

***What do you think are the main advantages and disadvantages of case-control studies? List your answer.***

*Advantages:*.....

.....

.....

.....

.....

.....

.....

*Disadvantages:*.....

.....

.....

.....

.....

.....

.....

*Exercise 3*

*List the advantages and disadvantages of cohort studies.*

*Advantages:*.....

.....

.....

.....

.....

.....

.....

*Disadvantages:* .....

.....

.....

.....

.....

.....

Answer 1

A Purpose: to investigate the incidence of diarrhoeal disease. Study design: descriptive, longitudinal.

B Purpose: to assess the effect of malaria on growth and development. Study design: intervention study.

C Purpose: to determine the prevalence of schistosomiasis. Study design: descriptive, cross-sectional.

D Purpose: to find out if raw fish was the source of the cholera outbreak. Study design: analytical, case-control.

Answer 2

Advantages of case-control (retrospective) studies:

- The study design guarantees the number of persons with the disease because we begin the study by selecting a group of cases.
- The study can be completed in a relatively short time (compared with prospective studies which may involve following patients for several years).
- Relatively simple to carry out.
- Relatively inexpensive.

Disadvantages of case-control studies:

- It is difficult to select the proper controls (see section 3.4).
- Determining exposure will often rely on the memory of the persons in the study and may be biased by knowing whether the person is a case or control.
- We cannot always be certain that exposure to the determinant occurred before the disease manifested itself.
- Persons who die as a result of disease caused by the determinant may not be known to the study.

Answer 3

Advantages of cohort (prospective) studies:

- They provide a direct estimate of the relative risk of an outcome associated with a particular exposure or factor, compared with case-control studies in which only an indirect estimate of relative risk, the odds ratio (see Unit 6), is possible.
- If strict criteria are defined, no selection bias should occur in selecting individuals for the cohort. The selection cannot be affected by the outcome as that is unknown at the start.
- We can be certain that the exposure or determinant was present before the disease outcome, and we can also note changes in exposure with time.

- We can study more than one outcome at a time, whereas in case-control studies the groups are defined by a single disease outcome.
- Data on persons who die will not be lost from the study providing that the cohort is selected at the start of exposure.

Disadvantages of cohort studies:

- If the disease in question is relatively rare, then this approach may require a large population and a long time, and thus would also be expensive.
- Participation in the study may influence the relationship between the determinant and development of disease.
- Problems may be created by people who drop out of the study.

---

## 3.4 WHOM DO I NEED TO STUDY?

### KEY POINTS

The important issues to be considered at this stage of planning an epidemiological study are summarised below. They will then be discussed in more detail in this section.

Study population
- Define physically and demographically the reference population on which information is to be sought (e.g. location, size, structure). This is needed in order to determine appropriate sampling procedures and eventual interpretation of the findings.

Sampling
- Decide whether the reference population is to be studied as a whole (comprehensive survey) or in part (sampled). This decision is based on the size of the reference population in relation to the resources available for the study. A comprehensive survey may in fact turn out to be a bad sample survey because of low response rates.

Sampling method
- If only a part of the population is to be examined, a method of sampling must be chosen which will ensure that those selected for study will be representative of the whole population. There are several scientific methods of selection: some are more practical than others. The best of samples can be ruined by low response rates.

Sample size
- Decide on the size of the sample to be selected. Optimum sample size depends on the prevalence or variability of the condition being surveyed, and the desired precision of the estimates. A sample size much larger than the optimum wastes resources. A sample much smaller than the optimum decreases the precision of the estimate and narrows the range of conclusions and generalisations which can be made.

Sampling bias
- Failure to sample appropriately may result in sampling bias, which reduces the extent to which it is valid to generalise from the sample back to the population.

Controls
- For analytical studies a control population must be selected for comparison with the study group. Controls must resemble the study group in all but the character under study. Observations on the study group and the controls must be made under the same conditions. If the controls are inappropriate the results of an analytical study are of little value.

### THE STUDY POPULATION

Descriptive studies

The population you choose to study will depend on the objectives of your investigation. For descriptive studies it is usually more convenient to confine the study to populations within defined geographical or administrative boundaries. For some kinds of study you may want to limit your investigations to populations with certain characteristics. For example, prevalence surveys for schistosomiasis

Special groups

may be limited to schoolchildren. However, take care when using special population groups, such as hospital or clinic populations, for descriptive surveys. The results which you obtain may not be representative of the population as a whole and need to be interpreted with care to avoid misleading conclusions.

Demographic information

In choosing your study population you will need some preliminary knowledge of the demography and residential pattern of the people. The demographic information necessary for planning must be obtained well in advance. Sometimes it can be found in offices but often it is only obtained by visiting the area and talking to the people who live there. Try to obtain a large-scale map of the area: you will find it invaluable both in planning your survey and in analysing your results.

Whereas in descriptive studies it is usual to define a study population and then make observations on a sample from it, in case-control studies observations may be made on a group of patients, the **study group**, who are not selected by formal sampling of a defined larger group. For example, a study on patients with snake bite may include every patient with this problem seen at one hospital during a certain period of time. For diseases such as schistosomiasis, which are common, formal sampling techniques (see below) may be used to select a representative group of cases from among all those known to have the disease.

## THE CONTROL POPULATION

Analytical studies

You will remember that in our discussion of analytical studies in section 3.3, we emphasised the necessity for a control population consisting of persons who were either unaffected by the disease (in case-control studies) or who were not exposed to the determinant (in cohort studies). A control group or cohort needs to be selected very carefully. The two general principles which govern the choice of controls are:

Principles

- The control group must resemble the study group in all but the characteristic which is under study.
- Observations made on the control group must be directly comparable to those made on the study group.

The comparison between cases and controls may reveal differences in exposure rates, and the development of disease may then be ascribed to this difference provided the two groups are otherwise comparable. In order to make sure that they are comparable, the two groups are frequently matched for characteristics that are known to influence the distribution of exposure, such as age or sex, and which are known as confounding variables.

Matching

Matching of cases and controls may be done either by stratification of the controls or by pairing. Stratification of a source of controls into age-groups, for example, and subsequent selection of different proportions of individuals from each strata, can be used to give a control group whose age distribution matches that of the cases. In pairing, a control is selected to pair with a given patient, being identical to the patient in respect of age and other confounding variables. Whichever method is used some degree of protection

against errors in matching is obtained by the use of several controls for each case.

Controls in case-control studies

In selecting individuals as controls the main requirements are:

- To achieve the necessary similarities between cases and controls in accordance with the principles defined above.
- To ensure that observations on the controls are made under the same conditions as those on the cases.

The possible sources from which controls can be selected for case-control studies include:

- hospital patients;
- relatives;
- neighbours;
- the total population.

Controls in cohort studies

The main requirements in the selection of controls for cohort studies are:

- The distribution of the controls must be the same as that of the study cohort in respect of any variable or attribute which influences the frequency of the disease (rather than in respect of any variable or attribute which influences both exposure of the determinant and the frequency of the disease, as in case-control stuides).
- Observations on the controls must be made under the same conditions as those on the study cohort.

The sources of control data in cohort studies are:

- a separate control cohort;
- individuals within the study cohort who do not share the same degree of exposure as other members of the cohort;
- national statistics.

Note the following points on using volunteers in an epidemiological study.

Volunteers

- Persons who volunteer to enter a study or submit to a procedure may differ in many respects from those who do not volunteer.
- In some circumstances, people who are anxious about their health may be those most likely to volunteer, in others they may be the most reluctant.
- Volunteers are usually more strongly motivated and more consistent in following instructions.
- However, the use of volunteers may be unavoidable when medical procedures or measurements are to be performed or evaluated, especially when multiple examinations are required over a long period of time.

Evaluating immunisation

It would be wrong to evaluate an immunisation procedure by immunising volunteers and then comparing them with persons who have not been immunised. A sounder procedure would be to carry out an experiment among volunteers by immunising some and not immunising others. Even then, the results of the experiment on volunteers, though well based for this group, may not necessarily be directly applicable to the population at large.

## SAMPLING

Once you have decided on the study population, you must next decide whether you will investigate the whole of the study population or only a sample of it.

Reasons for sampling

Often it is not possible, or even desirable, to study the entire population, as such a study would involve too much time, resources, manpower and money. Moreover the very size of the study might, in itself, introduce more errors into your findings. In some instances, however, examination of the entire population group is unavoidable; for example, when information is required on all cases during an epidemic, or when selection of a group of people for the study would create feelings of discrimination among the people.

Generalising from a sample

There is no difficulty in applying the results yielded by a sample to the parent population from which it has been selected, provided that certain conditions are met. Statistical techniques make it possible to state with what precision and confidence such inferences can be made (see Unit 7). The conditions to be met are:

Conditions for sampling

- The sample must be chosen so as to be **representative** of the parent population.
- The sample must be **sufficiently large**. If a number of representative samples drawn from the same parent population are investigated, it can be expected that, by chance, there will be differences between the findings in each sample; this problem of **sampling variation** is minimised if the sample is large.
- There must be **adequate coverage** of the sample. Unless information is in fact obtained about all, or almost all, of its members, the individuals studied may not be representative of the parent population, that is, there may be **sample bias**.

To make generalisations on the basis of sample results we need to consider how the sample relates to the parent population. The difference between the sample result and the population characteristic we seek to estimate is called the sampling error. There are two factors which contribute to sampling error:

Sampling error

Biased selection

- Biased selection - the sample is unrepresentative of the population. This is overcome by using a probability sampling method (see below).

Random variation

- Random variation - even if a sample is chosen in an unbiased way we would not expect it to be an exact replica of the parent population. The sampling error in this case is due to chance variation. This variation is determined by
  - the amount by which the characteristic varies in the population (e.g. ranges of blood pressures or haemoglobin values);
  - sample size (see page 93) since a small sample is a much less certain guide to the population from which it was drawn than is a large sample.

Refer to Unit 7 for more information on sampling in relation to estimating population values.

## SAMPLING METHODS

Probability sampling

The main method of sampling is 'probability sampling' in which each individual unit in the total population (each sampling unit) has a known probability of being selected. Generalisations can be made to the parent population with a measurable precision and confidence. We will discuss four types of probability sampling: random, systematic, stratified and cluster sampling. Probability sampling may be performed in more than one stage (two-stage or multi-stage sampling).

Simple random sampling
Method

Simple random sampling is a technique whereby each unit has the same probability of being selected. The basic procedure is:

- Prepare a 'sampling frame'. This is usually a list showing all the units from which the sample is to be selected, arranged in any order (e.g. a list of all the houses or people in a population).
- Decide on the size of the sample (see page 93).
- Select the required number of units at random by drawing lots or, more conveniently, by using a table of random numbers (see Appendix 1).

The ratio $\frac{\text{number of units in sample}}{\text{number of units in sampling frame}}$ is referred to as the sampling ratio or sampling fraction. It is usually expressed either in the form 1 in $n$, (for example 1 in 3, 1 in 4, etc.) or as a percentage or proportion.

Disadvantage

The main disadvantage of random sampling is that it is usually not practical to prepare a sampling frame which lists all individuals from whom the selection will be made. Preparation of such a list, would be too laborious and time consuming for the purposes of a single survey. One of the methods described below would be more appropriate.

Systematic sampling

This technique is often simpler than simple random sampling. The steps are as follows:

- Make a list (not necessarily numbered) of all the sampling units.
- Decide on the required sample size.
- Calculate the sampling ratio, expressed as 1 in $n$.
- Round '$n$' off to the nearest whole number and use this figure, $k$, as the sampling interval.
- Select every $k$th item on the list, starting with an item selected at random. For example, if every third person on a register is being selected then a random procedure must be used to determine whether the first, second or third person on the register is chosen as the first member of the sample.
  There are three possible samples:
  patients 1, 4, 7.....
  patients 2, 5, 8.....
  patients 3, 6, 9.....
  The sample obtained by this method can be considered as essentially equivalent to a random sample, providing that the list is not arranged according to some system or cyclical pattern.

Cluster sampling

In cluster sampling, a simple random sample is selected not of individual subjects, but of groups or clusters of individuals, and the sampling frame is a list of these clusters. The clusters may be villages, housing units, families, classes of school children, etc.

Advantages

The advantages of cluster sampling include:

- There is no need to obtain information to prepare a detailed sampling frame. Less detailed information, such as number of villages, only are required and this is often available.
- The cost of field investigations is reduced as it is easier and faster to obtain information on a group of people who are concentrated in an area, than to obtain the same information on people more widely scattered.
- Records of certain types of information are more likely to be accurate. For example, it might be easier to obtain the total number of deaths in a village than in a household.
- This technique is often more acceptable to the local community, as the entire community is examined. If only a few randomly chosen villagers were examined, they might become the objects of envy or pity.

Disadvantages

The disadvantages of cluster sampling include:

- If clusters contain similar persons it is difficult to estimate the precision with which generalisations may be made to the parent population.
- Cluster sampling may cause errors if the disease, attribute or variable being studied is itself clustered in the population. If, for example, a village chosen for the study of malaria is situated in an area with a high *Anopheles* population and this feature is not characteristic of the entire district, a survey of the village would then be likely to produce higher prevalence figures than would be present in the whole district.

Stratified sampling

To use stratified sampling, the population (sampling frame) is first divided into sub-groups or strata according to one or more characteristics, for example sex and age groups. Random or systematic sampling is then performed separately for each stratum. The end result is the selection of a sample that resembles the entire population more closely than one selected by simple random sampling alone.

Multi-stage sampling

This involves the selection of a sample in two or more stages. The stages are as follows:

- Set up a sampling frame of clusters to be used as primary sampling units (for example, villages or schools).
- Take a random sampling of these primary sampling units (for example a random selection of villages).
- Sub-divide this selected group into secondary sampling units (for example, houses).
- The secondary sampling units can be further subdivided if necessary.

Proper sampling techniques must be used to select the units at each stage.

Advantages

The main advantage of this sampling method is that fewer data are required than in preparing a simple random sample. Although complete listing of the primary sampling units would be required, information on the secondary sampling units is necessary only for the selected primary sampling units.

**Sample size**

Observations are made on a sample with the purpose of generalising from the sample to the entire population. The precision with which you can generalise from the sample is related to sample size. The smaller the sample the lower will be the cost, time and resources required for the study. Also it will be easier to minimise non-response or non-participation and maximise the accuracy and reliability of the information collected. However, if a sample is too small it may be impossible to make sufficiently precise and confident generalisations about the situation in the parent population, or to obtain statistical significance (see Unit 7) when associations are tested. On the other hand, it is a waste of time and resources to study a sample larger than is required to achieve the study objectives. How do you decide what sample size is needed for your study?

Small samples

Choosing sample size

In most descriptive community studies, the dominant considerations in the choice of sample size are practical ones - availablilty of time, staff and money. If you need to determine sample size and the degree of precision with which measurements in samples of a given size estimate the values in the study population, you can use reference tables or simple statistical formulae. These formulae differ according to whether the observations made are qualitative or quantitative.

**Qualitative data**

In qualitative (or quantal) data, only the presence or absence of a characteristic is recorded, with no estimate of magnitude. For example, presence or absence of *Ascaris* eggs. In order to estimate sample size you will need to:

Estimating sample size

- Estimate the proportion of persons with the characteristic under study: for example, estimate the prevalence rate.
- Decide the amount of sampling error that you are prepared to tolerate in the estimate rate (i.e. what is the desired degree of precision?).

Reference tables

Reference tables are available which show the minimum sample sizes required for various levels of expected prevalence and accepted margin of sampling error (precision) of the estimated prevalence. Figure 2 shows one such reference table.

[1]Figure 2 Reference table of minimum sample size for a prevalence survey

| Margin of error tolerated | Maximum expected prevalence rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1% | 2.5% | 5% | 10% | 20% | 30% | 40% | 50% |
| 0.5% | 1,522 | 3,746 | 7,300 | 13,830 | - | - | - | - |
| 1% | 381 | 937 | 1,825 | 3,458 | 6,147 | 8,068 | 9,220 | 9,604 |
| 2% | - | 235 | 457 | 865 | 1,537 | 2,017 | 2,305 | 2,401 |
| 5% | - | - | 73 | 139 | 246 | 323 | 369 | 385 |
| 10% | - | - | - | 35 | 62 | 81 | 93 | 97 |
| 15% | - | - | - | - | 28 | 36 | 41 | 43 |

Use

To use this table, we first select the appropriate column in the table according to how close to 50% the prevalence could possibly be or is ever likely to be. (If the figure is higher than 50% then use 100 minus the figure). Then we select the appropriate row in the table according to the amount of error due to sampling that we are prepared to tolerate in the estimated rate.

Figure 3 is another type of reference table. It shows 95% confidence intervals in different sample sizes, for a range of observed sample percentages from 5 to 95%. A 95% confidence interval gives the range within which we can be 95% confident that the true percentage or prevalence rate lies (see also Unit 7).

[1]From *Epidemiology for the health officer: a field manual for the tropics*. Edited by W. O. Phoon. WHO, 1985.

[2]Figure 3 95% confidence interval in different sample sizes where the observed sample percentage is in a range from 5% to 95%

| Sample size | Observed sample percentage | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 95% |
| 50 | 1-16 | 3-22 | 10-34 | 18-45 | 26-55 | 36-64 | 45-74 | 55-82 | 66-90 | 78-97 | 84-99 |
| 60 | 1-14 | 3-20 | 11-32 | 19-43 | 28-54 | 37-63 | 46-72 | 57-81 | 68-89 | 80-97 | 86-99 |
| 80 | 1-12 | 4-19 | 12-30 | 20-41 | 29-51 | 39-61 | 49-71 | 59-80 | 70-88 | 81-96 | 88-99 |
| 100 | 2-11 | 5-18 | 13-29 | 21-40 | 30-50 | 40-60 | 50-70 | 60-79 | 71-87 | 82-95 | 89-98 |
| 200 | 2-9 | 6-14 | 16-26 | 24-38 | 33-47 | 43-57 | 53-67 | 62-76 | 74-84 | 86-94 | 91-98 |
| 300 | 3-8 | 7-14 | 16-25 | 25-36 | 35-46 | 44-56 | 54-65 | 64-75 | 75-84 | 86-93 | 92-97 |
| 400 | 3-8 | 7-13 | 16-24 | 26-35 | 35-45 | 45-55 | 55-65 | 65-74 | 76-84 | 87-93 | 92-97 |
| 500 | 3-7 | 8-13 | 17-24 | 26-34 | 36-44 | 46-54 | 56-64 | 66-74 | 76-83 | 87-92 | 93-97 |
| 1000 | 4-7 | 8-12 | 18-23 | 27-33 | 37-43 | 47-53 | 57-63 | 67-73 | 77-82 | 88-92 | 93-96 |

Examples

Imagine that you examined 200 children for malaria and 30% were found to have positive blood slides, then there would be a 95% probability that the true positivity rate lay between 24% and 38%.

Alternatively, if you can estimate the prevalence and decide on a margin of error, then you can use this table to estimate the sample size. For example, if you estimate that the prevalence of malnutrition is 40% and decide that a 5% margin of error is acceptable (between 35% and 45% in this table), then you will need to measure 400 children.

*Exercise 4*

*You suspect that the prevalence of schistosomiasis is somewhere between 20% and 40% in your population and you wish to carry out a survey to estimate the actual prevalence with an accuracy of ± 2%. Using the table in Figure 2 determine the minimum sample size required.*

.....

.....

[2]From Swaroop, S. *Statistical methods in malaria eradication.* WHO, 1966.

*Exercise 5*

*Use Figure 3 to calculate the confidence intervals for the following sample sizes. In each case the sample population contains 30% with signs of leprosy.*

| *Sample size* | *Confidence intervals* |
|---|---|
| *50* | |
| *100* | |
| *500* | |
| *1000* | |

*Based on your results, what do you think would be the best sample size to choose. Give your reasons.*

.....

.....

.....

.....

.....

Answer 4

You will need to examine a random sample of at least 2305 persons. When you have completed your survey, if the sample shows a prevalence of 32.5% you can then be confident that the prevalence in your population (from which the sample was randomly drawn) is 32.5 $\pm$ 2%, that is between 30.5% and 34.5%.

Answer 5

| Sample size | Confidence intervals |
|---|---|
| 50 | 18-45 |
| 100 | 21-40 |
| 500 | 26-34 |
| 1000 | 27-33 |

If you choose a sample size of 50, the confidence interval suggests that the 'true value' for the entire population will be somewhere between 18 and 45%. Such a huge range gives no indication of whether leprosy is a minor or serious problem in the population. A sample of 500 individuals would indicate that the true prevalence for the population lies between 26 and 34%, an acceptably small and useful range, and accurate for most community studies. Sampling 1000 individuals would require more time and resources, but the results would be hardly less precise than if 500 persons were examined.

Confidence limits

Instead of using a reference table, the lower and upper limits of the confidence interval, known as confidence limits, can be calculated using the formula:

$$2 \times \sqrt{\frac{\%\text{ positive} \times \%\text{ negative}}{\text{number in sample}}}$$

Example

Let us see how this works using the example above.

| Sample size | Confidence limit |
|---|---|
| 50 | $2 \times \sqrt{\frac{30 \times 70}{50}} = 12.96$ |
| 100 | $2 \times \sqrt{\frac{30 \times 70}{100}} = 9.17$ |
| 500 | $2 \times \sqrt{\frac{30 \times 70}{500}} = 4.10$ |
| 1000 | $2 \times \sqrt{\frac{30 \times 70}{1000}} = 2.90$ |

If you choose a sample size of 100 the confidence limit of 9.17 suggests that the 'true value' for the entire population will lie somewhere between (30 - 9.17)% and (30 + 9.17)%, that is from 20.83% to 39.17%.

**Quantitative data**

In quantitative data measurements are made on a scale, for example, haemoglobin concentrations in g/dl. Such measurements will show variation among members of a population. This variation has two components:

Variation

- Differences in values between one individual and another.
- Differences in values in one individual at different times.

Both these differences may be affected by errors in the measurement technique (see Section 3.5).

For quantitative data the precision of sample estimates is expressed by the standard error of the mean and depends upon:

Standard error

- Variation in the measurements made. This variation is expressed by the standard deviation of the measurements (see Unit 5).
- The sample size.

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{\text{No. individuals in sample}}}$$

Estimating sample size

We can use this expression to derive an estimate of sample size. Imagine that we are sampling for the mean (average) level of haemoglobin in a population. We would need to state the following in order to calculate the minimum sample size we need.

Calculation

- How precisely we wish to estimate this mean level, i.e. the absolute amount of sampling error that we can tolerate (call this d).

- The standard deviation of the distribution of haemoglobin in the population (call this s). This can be obtained from published work or from a preliminary investigation.
- The chance that we are willing to take that we will get an unlucky sample giving a sampling error greater than d. Usually a 5% chance of error (giving a 95% confidence interval) is conventionally chosen.

The mean haemoglobin value $\pm$ d are the required 95% confidence limits, so that d = 2 x standard error of the mean (SE), where

$$SE = \frac{s}{\sqrt{n}} \qquad \text{Hence } \frac{d}{2} = \frac{s}{n}$$

Sample size (n) can then be estimated by

$$n = 2^2 \times \frac{s^2}{d^2}$$

Example

Let us work through an example. A health officer wishes to measure mean haemoglobin level in the community. From preliminary contact he thinks this mean is about 150 mg/l. He is willing to tolerate a sampling error up to 5 mg/l in his estimate. How many subjects will he need for his study?

If we assume the parent population is large, then the required minimum sample size can be calculated as follows. In this example s = 32 and d = 5.

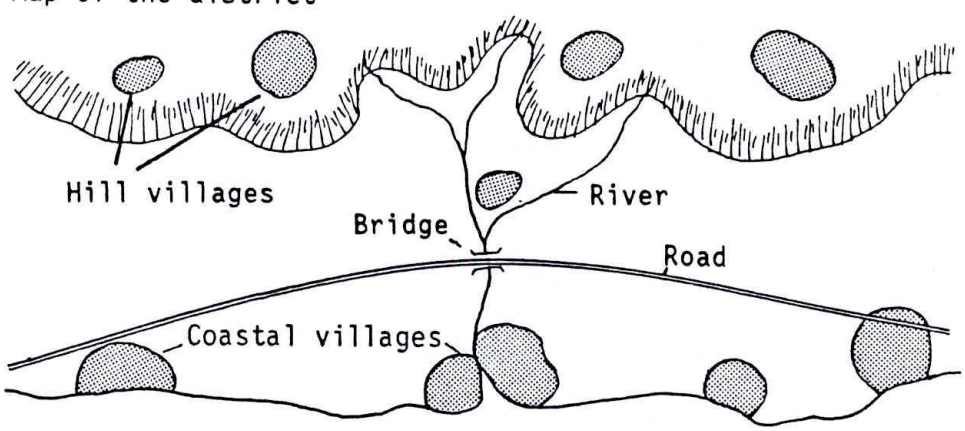
$$n = 22 \times \frac{32^2}{5^2} = 163.8$$

Therefore, the study needs at least 164 persons.

*Exercise 6*

*You want to measure the prevalence of a certain disease in an area with a population of 10,000. Obviously you cannot examine everyone. Look at the map (Figure 4) showing the distribution of the villages in the area. What do you think is a practical way of obtaining a representative sample of this population? Give reasons for your choice.*

Figure 4 Map of the district




.....

.....

.....

.....

.....

.....

*Exercise 7*

*What factors would you consider when choosing a sampling method for your investigation?*

.....

.....

.....

.....

.....

.....

.....

*Exercise 8*

*Make a list of the potential sources of error or bias that might occur in the composition or selection of study groups.*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Answer 6

If you choose to study just one village then your results will be biased because the area has both mountain and coastal villages and these might have quite different disease rates. The ideal situation would be to measure a few randomly selected people from each village but this is unrealistic because of the distances to be travelled and the expenditure this would involve. The hill villages may only be visited by walking so it would take a very long time to visit each village. The best practical solution is to first randomly select one coastal village and one hill village, and then within each village to randomly select the number of people to be examined.

Answer 7

The choice of a particular sampling method is largely determined by practical considerations. In fieldwork it is almost impossible to select a sample that is fully random. Every sample is therefore biased to some extent and in order to assess this bias the investigator must have some basic knowledge of the study population.

Answer 8

**Selection bias**

Selection bias may impair the validity of the findings as a measure of whatever it is that you wish to know about the general population, such as the prevalence of disease or the strength of an association with a disease. This is known as the validity of the study and should not be confused with the validity of measurements which we will discuss in section 3.5. Selection bias may arise from the following sources:

- Failure to choose a representative sample (sample bias), due to inadequate sampling methods or sample size.
- Shortcomings in the way that cases or controls were chosen (controls not comparable to cases).
- Incomplete coverage - failure to obtain information about all members of a sample of the population (non-response bias), refusal to participate in a study (non-participant bias) or the loss of members of the study population during an experiment or longitudinal survey because of migration, death, etc. (dropout bias).
- The whole study population may be a 'special' or 'different' one. This may result from selective admission to the population as in a study of hospital patients (admission rate bias) or groups which are characterised by their occupation or behaviour (membership bias). It may also result from selective migration into or out of a population. The 'special' nature of a study population of course does not lead to bias if you are only interested in the population which is studied, but generalisations which go beyond this population may not be valid.

---

## 3.5 WHAT OBSERVATIONS DO I NEED TO MAKE?

### KEY POINTS

The important issues to be considered at this stage of planning an epidemiological study are summarised below. They will then be discussed in more detail in this section.

Information required
- Determine the items of information required (i.e. select the variables to be measured). Only include those items necessary for the study to achieve its objectives (i.e. those which will be used in the analysis).

Definitions and criteria
- Specify working definitions for variables and standardised criteria for measurment and classification, and methods of collection. Make sure these are practical under the circumstances of the study.

Methods
- Select a method for measuring the variables, taking into account the reliability and validity of a measurement. Any method must be standardised to ensure that the results obtained in the study can be compared with those obtained in other studies.

Reliability
- A measurement is reliable when repetition of it gives the same results. A lack of reliability implies excess variation in the measurements. This variation can arise from variations in the characteristic being measured, the measuring instruments and the persons collecting the information (observer variation).

Validity
- The validity of a test refers to the extent to which it measures what it is intended to measure. Consideration of validity is an important aspect of choosing a diagnostic test, since it indicates the extent to which it is capable of correctly determining the presence or absence of the condition concerned.

### DECIDING WHAT OBSERVATIONS ARE NEEDED

Variables

The characteristics that are measured in an epidemiological study are referred to as variables.

Definition

Variables are characteristics which are measured either numerically (for example, age or height) or in categories (for example, sex or the presence or absence of disease).

Standardisation

Each of the variables measured in a study should be clearly and explicitly defined. Unless this is done there can be no guarantee that similar findings would be obtained if the study were repeated by the same or a different person. Therefore, it is important that all measurements should be reproducible, and you should spend considerable time during the planning stage of your survey to ensure that standardised criteria are used in measuring or categorising all variables.

Selection of variables

How do you decide which variables to include in your study? Selection of variables is based on the objectives of the investigation. If you have formulated your study objectives in writing, as we recommended earlier (see page 78), then you will have specifically

mentioned the key variables in the objectives. The more specific your objectives, the more variables will be included.

Relate to analysis

Do not collect information on variables which are not going to be used in your subsequent analysis of data. Be quite clear in your mind what information you hope to obtain before including a particular variable, because collecting and processing the extra data entails considerable effort and manpower. On the other hand, it is a mistake to limit the variables to such a small number that, at the end of the survey, you find it is not possible to analyse certain aspects of your study because no relevant information was collected. As a routine, therefore, you should first produce a comprehensive list of all the variables that you think might be required for the study. Go through each item in detail, to see how the information you wish to collect for that particular variable may be used in your ultimate data analysis. You may find it helpful at this stage to draw up the dummy (skeleton) tables you will use when you sort and summarise the data you have collected (see Unit 5). This will also help you to define how the variables should be classified, for example: do you need to record exact age or will you use age groupings - if so which age groups?

List of variables

How many?

The answer to the question 'How many variables should be studied?' is 'As many as necessary and as few as possible'.

Defining and measuring variables

Once you have decided upon the variables you need, the next step is to plan how to measure them under field conditions. In order to ensure that the observations are reproducible, you must decide on two things for every variable:

- Definition of the variable.
- Method for measuring the variable.

**Defining the variables**

What one person might call a 'common cold' another person might call 'influenza'. The same term may have more than one meaning, or mean different things to different people. Such differences in perception can lead to situations where the measurement of variables by different people will produce different results; that is, the findings are not reproducible. You must therefore define all variables clearly, and by a method that permits the variable to be objectively measured. There are two kinds of definition: conceptual and operational.

Conceptual definitions

The conceptual definition is equivalent to a dictionary definition. For example, the conceptual definition of malaria might be the presence of *Plasmodium* parasites in the patient.

Operational definitions

The operational definition (or working definition) defines the characteristic we will actually measure. It is phrased in terms of objectively observable facts and is sufficiently clear and explicit to avoid ambiguity. Where necessary it also states the method by which each of the facts are to be obtained.

An operational definition of malaria might be 'The presence of *Plasmodium* in the bloodstream of a patient as identified from a single thick blood film', or as 'A child with splenomegaly', or as 'A fever with chills', or as a combination of these.

Practicability

In formulating your operational definition of variables to be studied, you must always keep in mind that only simple, limited standardised techniques can be applied on a mass scale. More detailed examination techniques, such as those that are available in hospitals, are often not practical when large numbers of people need to be examined in places where there are no supporting facilities and when there is a tight schedule. It has to be accepted that the employment of such simplified techniques might lead to missing a percentage of 'cases'. It is important to know or estimate what this proportion might be (see page 113 for information on choosing a diagnostic test). You must also ensure that your findings are reliable (see page 104).

**Defining disease and diagnostic criteria**

Accurate diagnosis

Few diseases have satisfactory and widely accepted working definitions. However, accurate diagnosis is as important to you as an epidemiologist as it is to you as a clinician. As a clinician your task is to decide to which of many diagnostic categories your patient belongs. You have to answer the question, 'What condition does this patient have?' You are free to perform as many additional studies as needed until the diagnosis becomes clear. By contrast, as an epidemiologist you have to preselect the diagnostic criteria so that you can answer the question, 'Does this individual in my population sample have the condition I am studying or not?'

Specificity of criteria

If you need to identify all persons who almost certainly have the disease, you will need a highly specific definition, using criteria which may fail to identify many people who clinicians say have the disease. On the other hand, if your aim is to detect all persons who may possibly have the disease, even at the expense of falsely including many who do not have it, you can use less stringent criteria.

Choosing criteria

The choice of diagnostic criteria to be used is heavily influenced by the methods by which data are to be collected. Very different criteria may be used in a study based solely on history taking, one in which a clinical examination is done, or one in which biochemical, micro-biological, radiological and other diagnostic tests are used. Consider two extreme definitions of chronic bronchitis.

Example

A definition specifying 'chronic inflammatory, fibrotic and atrophic changes in the airways' is obviously of no diagnostic use on living patients. At the other extreme, no medical training is required to diagnose chronic bronchitis with the use of a standard questionnaire if it is defined as 'the production of sputum from the chest at least twice a day on most days for at least three months each year for two or more years'.

New cases

If you are studying the incidence of disease, you will need to define the meaning of a 'new case'. For example, after what period of freedom from symptoms would a patient be regarded as having a new episode of acute bronchitis?

Date of onset

You should also record the date of disease onset and define whether this refers to the date when symptoms were first noticed or when the disease was diagnosed or when the disease was first notified.

Selecting a test

In selecting the diagnostic test and the criteria to be applied, you need to consider the levels of diagnostic accuracy and the predictive value of the different methods that are available (see below, page 107).

**Measuring the variables**

When you choose the methods for measuring the variables, you need to consider two aspects:

- The reliability of the measurement.
- The validity of the measurement.

Let us first explain what we mean by reliability and validity.

Reliability

Reliability: a result of measurement is said to be reliable when it is stable, that is, when repetition of an experiment or measurement gives the same results.

Repeatability, reproducibility and precision are aspects of reliability. A lack of reliability implies excessive variation.

Variability in a measurement has a number of sources:

- Measurement
  - Instrument (the means)
  - Observer (the person).
- Biological
  - Within individuals (changes with time and situation)
  - Among individuals (biological differences from subject to subject).

Validity

The validity of a measurement refers to the the extent to which it measures the characteristic that the investigator actually wishes to measure. Another word for validity is accuracy.

**RELIABILITY**

Read through the following examples which illustrate the concept of reliability.

- The incidence of accidental injuries was studied in a rural area by paying periodic home visits and asking about injuries occurring since the previous home visit. The incidence was doubled when the inquiries were made at intervals of two weeks instead of a month.
- Two highly skilled ophthalmologists examined the same population sample for evidence of trachoma, using the same diagnostic criteria (physical signs) and examination procedure. One found 125 cases of active trachoma and the other found 138, but these included only 77 who were diagnosed by both experts; the other 109 were diagnosed by only one or other of them.
- The same chest X-ray films were examined independently by five experts in order to determine the presence of tuberculosis. Although 131 films were recorded as positive by at least one expert, there were only 27 about which all five experts agreed. The films were later looked at by the same observers, and there were many reversals of verdict. For example, one radiologist who had found 59 positive cases the first time, found 78 the

second time, including 55 who had been positive the first time and 23 new cases.

- Material with a known concentration of haemoglobin (9.8 g/dl) was sent to a number of hospital laboratories, and a separate determination of haemoglobin was made in each laboratory. The results ranged from 8 to 15.5 g/dl.
- A standard suspension of white blood cells was examined in a number of laboratories. When visual counting was performed, the white blood cell counts ranged from 3.7 to 6.9 x $10^3$ cells/$mm^2$; when electronic cell counters were used the range was from 2.4 to 7.1 x $10^3$ cells/$mm^2$.
- A number of studies have shown that if children are measured in the morning they are on average 0.5 cm or more taller than if they are measured in the afternoon.

Now try to answer the following question.

*Question*

*What do you think are the main sources of variation between measurements?*

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Variation between measurements

Variation between measurements has three sources.

1 Variation in the characteristic being measured may be caused by variation in any of a whole complex of factors which determine the characteristic. These factors will include the measuring procedure itself. For example, the response to a question may be affected by the way the interviewer asks the question, his appearance, sex or manner. A subject may change his behaviour if he knows he is being studied.
2 The measuring instruments may not provide consistent results, or different instruments may give different results. Note that the term 'measuring instruments' refers not only to mechanical devices, but biochemical and other tests, questions and questionnaires.
3 The persons collecting the information, including interviewers and people carrying out examinations or tests, may vary in what they record, in their skill, commitment, inclination to make mistakes, etc. They may be influenced, either consciously or unconsciously, by their own expectations and motivations.

Observer variation

This source of variation is known as 'observer variation', because it arises from the persons making the observations and not from changes in the characteristic being measured or from the measuring instrument. There are two kinds of observer variation or error. Between-observer variation arises if one observer examines a blood film and diagnoses malaria, while a second observer variation arises if a single observer performs two consecutive measurements of arm circumference in the same patient and obtains two different results.

Measuring reliability

Although you should obviously make an effort to collect reliable information, it is not essential for you to aim for complete reliability. It is important, however, to know how much unreliability there is, particularly with regard to the variables which play an important part in the investigation.

For each of the three sources of variation identified above, a general way of testing reliability would be:

1 Assessment of variation in the characteristic being measured must depend on repeated measurements in the same subject.
2 The limits of accuracy of a measuring instrument and the extent to which any one shows consistent differences from others is defined by comparison of measurements made with different instruments.
3 Assessment of within-observer variation requires repeated measurements of the same phenomenon by one observer.

Reducing variation

You could try the following ways of reducing variation to reasonable limits and so enhance reliability.

- Variables should have clear operational definitions with clear and detailed descriptions of the methods by which the information is to be collected.
- A standard procedure of examination should be used, and standard questions asked in a standard way.

- If the procedure is one requiring special skill, the necessary training should be provided.
- If there is more than one observer, they should make sure that their methods are consistent, possibly by working together for a while to minimise between-observer variation.
- If a high degree of within-observer variation is found, either a single observer should collect all the data, or each individual should be seen independently by two observers.
- An instrument should be chosen which supplies relatively consistent measurements.
- The variation associated with the instrument should be small in relation to the total range of variation of the variable being measured.
- If more than one mechanical measuring device is used they should be of the same model and/or standardised against each other.
- Equipment should be tested from time to time and used under identical conditions.
- With quantitative measurements which show appreciable variation, the average of two or more readings may be used.

## VALIDITY

Definition

The validity of a measurement refers to the extent to which it measures the characteristic that the investigator actually wishes to measure. A routine of clinical examination to detect nutritional disorders, for example, must be constructed so that it detects as many people as possible who have the disorders while generally excluding those who are normal or who have other conditions.

Consideration of validity is an important aspect of choosing a diagnostic test or diagnostic criteria (see page 113). By diagnostic test we mean any method of measuring variables, such as a physical examination, laboratory test or questionnaire.

Accuracy of a test

The accuracy of a diagnostic test refers to the extent to which the test is capable of correctly diagnosing the presence or absence of the disease concerned. When a test is performed, some people will be wrongly measured as positive (false positives) and some people will be missed (false negatives). Comparison of the number of false positives and false negatives gives two measures of the validity of a test: sensitivity and specificity.

Sensitivity
Specificity

Sensitivity is the ability of a test to correctly identify those who have the disease; specificity is the ability to identify correctly those who do not have the disease. For example, a test has a sensitivity of 80% if it correctly gives a positive result in 80% of persons who actually have the disease. A test has a specificity of 80% if it correctly gives a negative result in 80% of persons who actually do not have the disease.

Evaluation

Diagnostic tests are generally judged on the basis of their sensitivity and specificity. Such evaluations are essential but they may not provide all the information that a user of a test may need to make decisions concerning the best diagnostic strategy for the particular circumstances of the study.

Predictive value

The predictive value of a test, which depends on the disease prevalence as well as the sensitivity

and specificity, is the most important measure for determining its usefulness under field conditions.

Example

Let us illustrate what we have just discussed using the example of childhood tuberculosis (see Figure 5). In the community there will be an overlap between the healthy population and the tuberculous population, consisting of those persons who have met the bacillus and have become sensitised but who have not developed the disease. The majority will be tuberculin positive but only a proportion will have tuberculosis. Tuberculin testing is therefore a sensitive test for active pulmonary tuberculosis, but not a specific one.

Chest X-ray is less sensitive than tuberculin testing, since a proportion of patients with active disease do not have radiological changes. But it is more specific than tuberculin testing because the proportion of patients with active disease among those with positive X-rays is higher than among those with positive tuberculin reactions.

The sputum test is the most specific (but not very sensitive) of all the tests and would therefore be useful as a screening test in community diagnosis where cost and applicability are important. The information presented in Figure 5 can be summarised as follows.

Figure 6 Test results by diagnosis

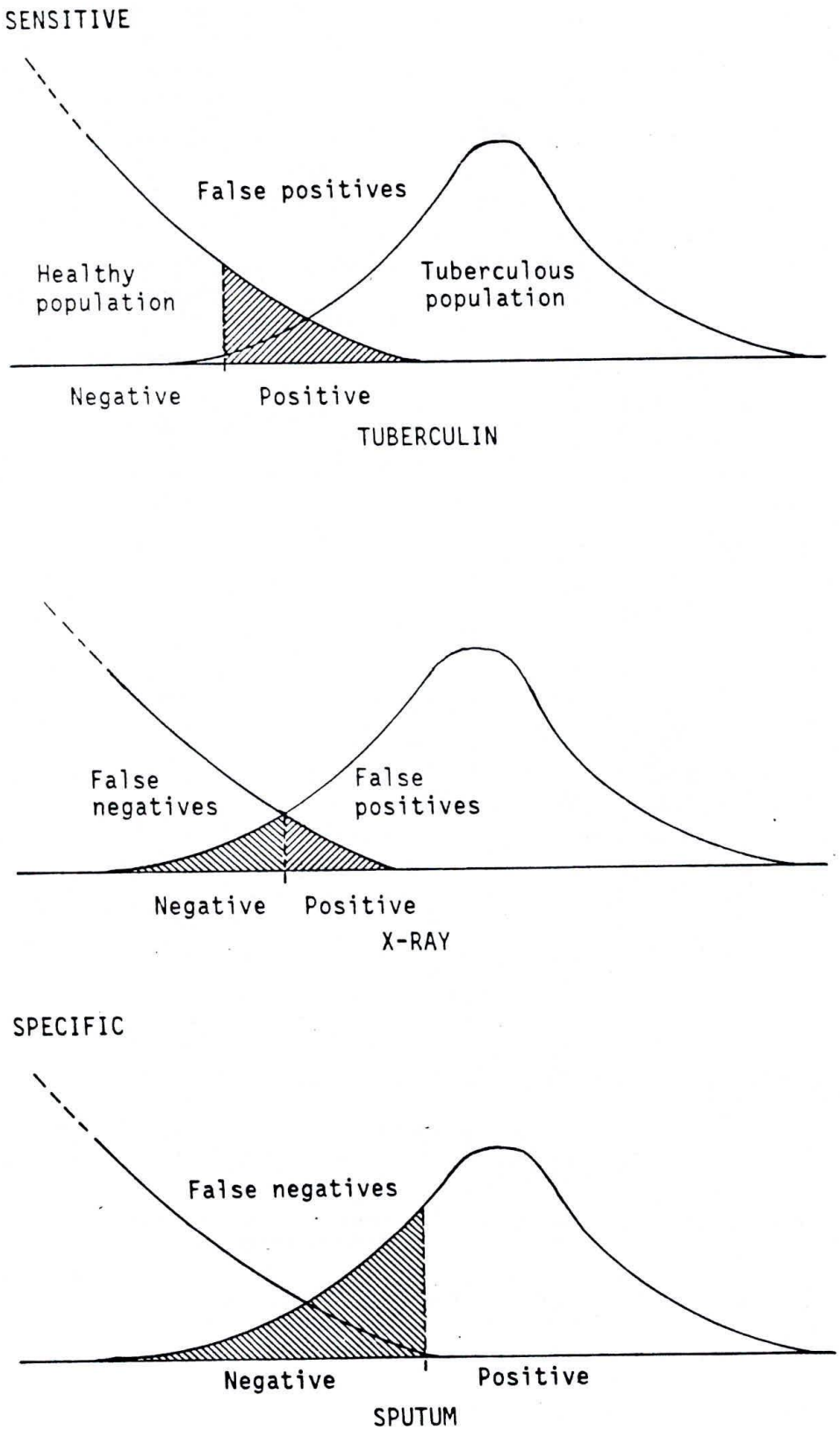
| Test results | Diseased | Not diseased | Total |
|---|---|---|---|
| Positive | a | b | a + b |
| Negative | c | d | c + d |
| Total | a + c | b + d | a + b + c + d |

a = diseased persons detected by the test (true positive)
b = non-diseased persons detected by the test (false positive)
c = diseased persons not detected by the test (false negative)
d = non-diseased persons negative to the test (true negative)

The following measures are used to evaluate a test:

| | | | |
|---|---|---|---|
| Sensitivity (True positive) | = | Percentage of persons with the disease who were positive to the test | $= \frac{a}{a+c} \times 100$ |
| Specificity (true negative) | = | Percentage of non-diseased persons who were negative to the test | $= \frac{d}{b+d} \times 100$ |
| False negative | = | Percentage of persons with the disease who were negative to the test | $= \frac{c}{a+c} \times 100$ |
| False positive | = | Percentage of persons without the disease who were positive to the test | $= \frac{b}{b+d} \times 100$ |
| Predictive value of a positive test | = | Percentage of persons with a positive test who have the disease | $= \frac{a}{a+b} \times 100$ |
| Predictive value of a negative test | = | Percentage of persons with a negative test who do not have the disease | $= \frac{d}{c+d} \times 100$ |

Figure 5 Relationship between negatives, positives, sensitivity and specificity using childhood tuberculosis as an example.



Look at the following test results obtained on a sample of 2000 persons before and after a disease control programme.

Before control: prevalence rate is 20% (400 out of 2000 have the disease)

| | | Disease + | Disease - | Total |
|---|---|---|---|---|
| Test result | + | 380 | 80 | 460 |
| | - | 20 | 1520 | 1540 |
| Total | | 400 | 1600 | 2000 |

After control: prevalence rate is 1% (20 out of 2000 have the disease)

| | | Disease + | Disease - | Total |
|---|---|---|---|---|
| Test result | + | 19 | 99 | 118 |
| | - | 1 | 1881 | 1882 |
| Total | | 20 | 1980 | 2000 |

*Exercise 9*

*Calculate the test sensitivity, specificity and predictive value (positive) before and after disease control. What are the implications of the answers you obtain?*

.....

.....

.....

.....

.....

.....

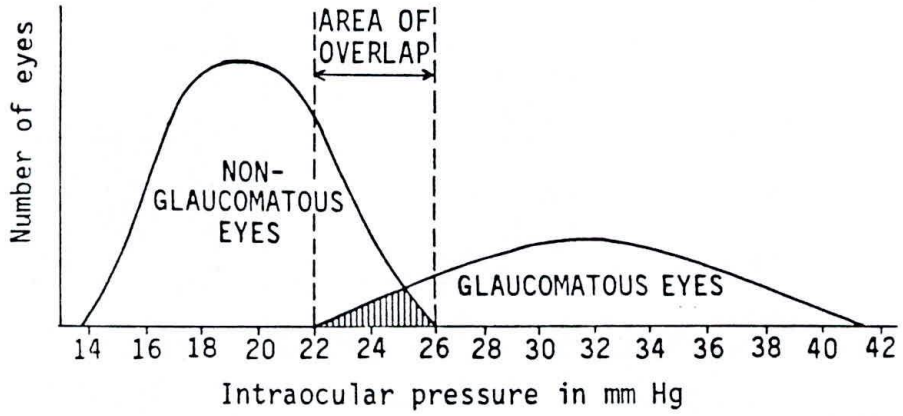
.....

.....

.....

Figure 7 depicts the eye pressure distributions for normal and diseased eyes in glaucoma.

Figure 7 Intraocular pressure of glaucomatous and non-glaucomatous eyes




*Exercise 10* *If the diagnostic criterion is set at 22 mm Hg, what can you say about the sensitivity and specificity of the test?*

.....

.....

.....

*Exercise 11* *If the diagnostic criterion is set at 26 mm Hg, how do the sensitivity and specificity change?*

.....

.....

.....

*Exercise 12* *What general principle does this example illustrate?*

.....

.....

.....

.....

.....

.....

*Exercise 13*

*If you were using this as a screening test for glaucoma, where would you establish the diagnostic criteria? Give your reasons.*

.....

.....

.....

.....

Answer 9

| | Before control | After control |
|---|---|---|
| Sensitivity | 380/400 = 95% | 19/20 = 95% |
| Specificity | 1520/1600 = 95% | 1881/1980 = 95% |
| Predictive value (+) | 380/460 = 83% | 19/118 = 16% |

Prevalence and predictive values

The calculations show that even for a test of high sensitivity and specificity the predictive value of its positive results falls dramatically from 83% to 16% when the prevalence of the disease falls from 20% to 1%. Before the control programme, nearly everyone who had a positive test actually had the disease; when the prevalence was reduced to 1% only about one of six whose tests were positive actually had the disease, yet there was no change in the sensitivity or specificity.

New tests

It is important to remember that predictive values (and proportions of false positives or false negatives) are dependent on the frequency of a disease in the study group whereas, in theory at least, sensitivity and specificity remain constant, irrespective of the frequency. New diagnostic and screening tests are often tried out on groups of hospital patients among whom the frequency of the disease being studied is high. The estimates of sensitivity and specificity obtained in these studies may be used in the planning of epidemiological studies in the general population, where the frequency of the disease is usually much lower. However, the predictive values and proportions of false positives and false negatives will not be the same in the hospital patients and the general population.

Answer 10

If the disease is diagnosed when examination reveals 22 mm Hg pressure, sensitivity will be 100% but specificity will be below 100% because a large number of healthy eyes will be falsely diagnosed as being glaucomatous since 22 mm Hg is within the range of normal intraocular pressures.

Answer 11

If the diagnostic criterion is raised to 26 mm Hg, specificity will be 100%, but sensitivity will be below 100% because those glaucomatous eyes which have an intraocular pressure between 22 and 26 mm Hg will not be diagnosed.

Answer 12

It is obvious that the sensitivity and specificity cannot both be 100% while the distributions of well and diseased populations overlap with respect to the variable being measured by the test. Sensitivity and specificity are usually inversely related and therefore one may be increased only at the expense of the other.

Answer 13

The point at which the distributions intersect is frequently used as the criterion because it will generally minimise the false positives and false negatives when both are considered important. The decision on the criterion used is generally a reasoned judgement about the number of false positives and false negatives that are tolerable to the population and the provider of the screening service. This judgement should be based on the severity of the disease, the cost of the test, the time taken to administer it, and the advantages and probability of obtaining early treatment.

## CHOOSING A TECHNIQUE

During the planning phase of the study you will need to decide upon the method to be used for collecting information concerning each of the variables listed for investigation. The choice of methods is largely based on the quality of the information they will yield: two aspects of this quality, reliability and validity have just been discussed.

Practical aspects

The selection of a method is also based on practical considerations, such as:

- The need for personnel, skills, time, equipment and other facilities in relation to what is available.
- The acceptability of the procedures to the study population.
- The probability that the method will provide a good coverage, that is, it will supply the required information about all or almost all the members of the population or sample. For example, it is not appropriate to ask a question that only a few people will be able to answer.

These practical aspects should be considered not only in relation to the measurement of each separate variable, but also in relation to the methods of data collection as a whole.

Collecting information

The methods of collecting information can be broadly classified as follows:

- Observation - the use of techniques varying from simple visual observations to those requiring special skills such as clinical examinations, or sophisticated equipment or facilities, such as radiographic, biochemical or microbiological examination.
- Interviews and questionnaires - see Unit 4.

- Documentary sources - clinical records and other personal records, death certificates, published mortality statistics, census publications, etc.

Choosing a method

Accuracy and practicability are often inversely correlated. A method providing more satisfactory information will often be a more elaborate, expensive or inconvenient one. However, remember that the aim is not 100% accuracy, but the optimal accuracy required for the purposes of the study, and consistent with practical possibilities. Information on the accuracy and practicability of the proposed methods can often be obtained from previous studies or from experiences of other investigators. Usually, however, you will find that you need to test at least some of the proposed methods in a small-scale pilot study.

Pilot study

For example, it may be necessary to determine how long an interview or examination may take, how acceptable it is, or whether questions are clearly intelligible and unambiguous. We shall consider these issues more fully in Unit 4.

---

## 3.6 PUTTING THEORY INTO PRACTICE

**KEY POINTS**

The main issues to be addressed at this stage of planning an epidemiological study are summarised below and then discussed in more detail in this section.

Resources
- Decide what resources (manpower, materials, finance, time) are needed in order to carry out the study.

Timing
- Decide when to carry out the study.

Location
- Decide exactly where to do the study.

Information
- Obtain detailed information on the study area. Meet the community leaders, tell them your plans and get their approval.

Formalities
- Complete any national or local formalities which are needed before beginning a study.

Analysis
- Make plans for analysing the data which will be collected.

Protocol
- Write a study protocol outlining the objectives and methods (what is to be done, by whom, when, where and how), timing and resources.

Introduction

So far in this unit we have considered:
- Definition of study objectives.
- Choice of type of study.
- Choice of study population and sampling technique.
- Definition and measurement of variables.

In particular, we have looked at the theoretical considerations underlying these aspects of planning, so that we can minimise the errors and variation, and ensure that our study is based on sound principles. Now we shall consider some of the more practical aspects of planning - tasks which need to be done, or decisions which need to be made before the plans can be translated into action.

**RESOURCES**

We suggested at the beginning of this unit (page 80) that when you formulate your study plans you should take into account the amount and type of resources which are available to you, so that your plans are realistic and practical. Your task now is to decide exactly what resources you need in order to carry out the study you have planned. The general requirements will be for manpower, materials, finance and time.

Manpower

An analysis of what needs doing, when, where and how often will help you decide on the type of personnel that you will need. You should make a job description for each post that is required and establish criteria for the selection of personnel. The number of staff required will vary with the size and complexity of the study and the

period of time over which the data will be collected. Unit 4 gives more information on selection and training of personnel.

Materials

It is a good idea to estimate the minimum materials that would be required for the study and then add a certain extra amount to allow for unforeseen developments or underestimation. It is always better to have a little more of the basic requirements for the study than to have too little, otherwise the field workers may be forced to use techniques that would not normally be acceptable in order to meet the scheduled targets.

Finance

Financial requirements are, of course, the basis of all other requirements. Two major items of expenditure will be transport costs and the payment of any extra personnel you need to hire to help you collect the data. Always allow a margin for unforeseen expenses.

Time

The links between the different stages of a study and the time needed for each stage to be completed will give you an indication of the time required to carry out the entire study.

### TIMING

An investigation should be implemented according to the target dates and completed within the time scheduled.

There are many factors which may influence the timing of your study. Some of these may be particular to your area and will depend on a thorough knowledge of local customs, festivals, culture and other social or occupational activities. However, the following general points can be made:

Rainy season

- Field surveys during rainy seasons may run into transportation and communication problems.

Seasonal activities

- Seasonal activities may affect the activity and mobility of the population. Surveys carried out at planting or harvesting periods may yield a poor response because most people would be working in the fields. Therefore the surveys may need to be carried out at the beginning or end of the day.

Markets

- Market days may be an advantage or a disadvantage, depending on the type of survey being carried out. Regularly held markets will affect population movements in the area and this should be considered when planning visits in village-based studies.

### LOCATION

By this stage you will have already selected the general area for your study and stated the basis for selecting the samples. Now you will need to select the exact area as, in practice, several constraints may make your predetermined selections impractical. For example, a village selected purely by random sample, may be virtually inaccessible if it can only be reached after a 25 kilometre walk.

Exact area

Accessibility
Acceptability

Any area which you select must be accessible. It is also essential that your investigation is acceptable to the local community. There are many kinds of local customs, beliefs and practices which can cause problems and lead to the rejection of some investigations. In

some communities, for example, collection and removal of faeces is not acceptable.

Preliminary visit

A good understanding of local customs and culture is therefore necessary, and it is therefore a good idea to make a preliminary visit to the study area. This helps not only to overcome existing problems but also to prevent problems from arising in the future through ignorance and insensitivity about local customs. Cultural practices may vary markedly even in a relatively small area, and obtaining information through written or verbal reports does not provide the same depth of understanding as a personal visit. Time spent in the area can also be used to build up a close rapport with the village leaders.

## FORMALITIES

Approval

Before putting your plan into action you should ensure that you have informed the appropriate authorities and obtained the necessary approval to do the study. In addition to national formalities there are usually local formalities to be completed; both will vary from country to country.

## PLANNING THE ANALYSIS

Skeleton tables

'Think forward'

During the planning phase you should decide, at least in broad outline, how the information you propose to collect will be analysed. It is often helpful to draw up a number of specimen skeleton tables (refer to Unit 5 for help with tabulation) showing the scales of classification of the variables they include (that is, with column and row headings but containing no figures), and to consider how different kinds of results will be interpreted. This process of 'thinking forward' to the analysis often reveals gaps in the data, defects in scales of measurement, or an excess of certain data. It provides a further opportunity for second thoughts as to whether the study as planned is likely to meet its objectives.

## WRITING THE PROTOCOL

Written protocol

Discussion

By the end of the planning phase you should have a written study protocol, outlining the study objectives, methods, timing and resources. This will help to clarify your thinking and remind you of what you have decided. It is also a good idea to discuss your plan with your colleagues and your supervisor (if you have one) before you start to implement your study. If, as we hope, you write a report at the end of your study, you will need to give an account of your objectives and methods. If you leave all the writing to the end you will have forgotten many details about the methods you used.

---

## 3.7 SUMMARY

In this unit we have considered many aspects of planning an epidemiological study. We can summarise these as a checklist of tasks which need to be done during the planning phase. Remember that many of these tasks are interdependent, and we are not suggesting that you should work through the list systematically. However, you should have considered all these points before you begin to translate your plans into action.

Planning checklist

- Define objectives and type of study to be carried out (sections 3.2 and 3.3).
- Determine how much money and manpower are available (sections 3.2 and 3.6).
- Look up reports of previous studies and consult people with experience in the particular field (section 3.1).
- Choose a study population and a method of sampling it (section 3.4).
- Choose a method of selecting controls if necessary (section 3.4).
- Define observations to be made and choose standardised techniques for making them (section 3.5).
- Make a preliminary appraisal of the area to obtain demographic, social and cultural data (section 3.6).
- Complete national or local formalities necessary to obtain permission for the survey (section 3.6).
- Decide timing of survey (section 3.6).

# UNIT 4: COLLECTING DATA

# UNIT 4: OBJECTIVES

Study with this unit will enable you to collect data for an epidemiological investigation in your district.

In particular you will be able to:

- Design and evaluate record forms, including questionnaires, precoded record forms and peripheral punch cards, for the collection of data.
- Apply the principles of organising an epidemiological study to the collection of data in your district.
- Train personnel to assist in the collection of data.
- Monitor and evaluate data collection.

# UNIT 4: CONTENTS

## 4.1 INTRODUCTION

This unit is a short one. It is also a very important one.

All the work done during the planning phase of a study has been a preparation for the collection of data; the subsequent analysis, interpretation and use of the results depend on the data which are collected. Therefore, data collection is the core of an epidemiological study and it is important to make sure that it is done:

- **accurately** - to ensure that the data are reliable and valid;
- **efficiently** - to ensure that the best use is made of your limited resources;
- **sensitively** - to ensure that you obtain and maintain the cooperation of the community you are studying and the staff who are collecting the data on your behalf.

This unit will help you to achieve these goals.

We begin the unit by considering exactly how the data can be recorded, and then go on to discuss the practical preparations which need to be made before data collection in the field begins. Finally, we will consider your role in monitoring the progress of the investigation during the period of data collection.

---

## 4.2 HOW DO I RECORD THE DATA?

Data recording cannot be planned until the study plan has been completed because you cannot plan the records until you know what variables will be studied, what scales of measurement will be used, and how the information will be collected and processed (see Unit 3, section 3.5).

Importance of design

Recording data is an essential part of any epidemiological study, and the ease and accuracy with which this is done depends partly on the design of the record form. A carefully designed form facilitates not only the collection of reliable data, but also its storage, sorting and analysis. These are the aspects of data collection we shall be considering in this section of the unit.

Two further points about data collection need emphasising, and these will be dealt with later in the unit. Firstly, it is usually the field worker who supplies the crude data. Therefore, it is most important that he is trained and motivated (see section 4.3).

Field workers

Accuracy

Secondly, the recording of data must be carefully checked for accuracy during the course of the study (see section 4.4).

### DESIGNING THE RECORD FORM

There are two basic types of data which could be collected in any investigation:

Questions

- Data obtained from the respondent or his family, usually from interviews and questionnaires.

Examinations

- Data obtained from examining the respondent, or specimens collected from him such as blood or faeces.

There are two types of form which can be used for the collection of information.

Self-administered questionnaires

- Self-administered questionnaires are forms which are given to the respondent to complete by himself. Obviously these are not suitable for obtaining information from people who are illiterate. You may find them useful if you wish to do a postal survey: for example, sending questionnaires to teachers to find out information about the children in their schools.

Recording forms

- Recording forms are forms which are filled in by an enumerator (a trained field worker) who obtains the information by interview and observation. It is important to ensure that the recording forms are standardised so that data collected by different workers are accurate and reliable. One recording form should be used for each unit being studied, such as a person or household. The recording forms will need to be processed to produce the information required, and it is important to design the form so that it can be easily sorted for analysis.

The main points to consider when designing a form are:

Content
- Ensure that you have included all the information which will enable you to achieve the study objectives or test the study hypotheses (see Unit 3, section 3.2). It is a good idea to begin by listing all the variables the form is designed to record, and then to formulate valid questions which will measure these variables (variables and validity were discussed in Unit 3, section 3.5). In addition to the information specific to the study, a record form for an individual must usually include age, sex, address and personal identity or name.

Length
- Avoid the use of very lengthy forms. Much time and energy are often wasted in the collection of information which has no relevance to the objectives of the survey. Furthermore, a lengthy form may be resented by the respondent who is under no obligation to participate in your study. If the respondents are not motivated to answer the questions fully, the information gained is unlikely to be accurate or complete.

Order
- Arrange the questions in order of difficulty. The first questions should be easy, and uncontroversial. It is often wise to leave any difficult or embarrassing questions until the end of the form, so that the interviewer has established a closer rapport with the respondent before he asks these questions.

Language
- Phrase the questions in clear and simple language. Avoid technical terms and check your questions for possible ambiguity. Remember that you are usually dealing with someone who is not as well educated as you are. Try to phrase the questions in such a way that they will sound as though the interviewer is having a conversation with the respondent rather than interrogating him.

Practical
- Make sure that the respondents can answer the questions. For example, there is little point asking about events or experiences which many subjects will not remember, such as minor illnesses or the food he ate several days ago.

Filter questions
- If some questions are not relevant to all the respondents, filter questions can be used which enable the interviewer to omit certain questions. For example, if a respondent answered 'Yes' to the question 'Has your child received any vaccinations?' the following questions may ask for further details about the vaccinations. If the respondent answers 'No' then the interviewer should be directed to miss out these irrelevant questions.

Layout
- Decide on the layout of the form. Allow enough space for the information to be written down on the form. Think about how you are going to sort and process the data when it has been collected. Information can be sorted more easily if the record forms are coded (see page 126). Information can be processed more easily if punched cards are used for recording (see page 129).

Example

Figure 1 shows a pre-coded questionnaire designed to assemble information about measles cases admitted to a hospital. We shall discuss coding in more detail shortly, but first note the following general points.

Title
- The form has a title so that the study for which the form is to be used can be identified.

Identification
- The form has an identification number which also includes useful information (month and year).

Instructions
- The instructions at the top of the form make it clear how the form is to be completed.

Name
- The respondent's name is included for identification and subsequent follow-up. If, for reasons of confidentiality, it is not appropriate to record names, then record the household number or house number for identification purposes.

Address
- The person's address should be requested according to local practice, for example village and district. Get the name of the village leader for follow-up.

Age
- Age is recorded onto a series of numbers. It could also be recorded within groups: the convention is to use five-year intervals such as 0-4, 5-9, 10-14, etc. If follow-up is necessary, date of birth can allow subsequent age to be calculated.

Question type
- Questions can be either 'open' or 'closed'. Closed questions are answered by choosing from a number of fixed alternatives (as in question 10), and are easier to analyse. The 'other' section in question 11 is left open in case an analysable group occurs that was not specifically included in the 'closed' section.

Closed questions

Open questions

Closed questions limit the information collected, but take little time for respondents to answer and are more easily coded for analysis (see below). Open questions enable a greater amount of information to be collected, and are particularly useful for questions relating to attitudes and opinions. However, they generally demand more time of respondents, and will be more difficult to code and analyse.

Remarks
- It is sometimes a good idea to include a section for 'remarks', especially on postal questionnaires, as this encourages people to express themselves fully.

**CODING**

Definition

Coding of information consists of the assigning of a numerical code to a specific item of information, that is, it simply means changing the information into a numeral.

For example, we may use the numerical value [1] to denote a 'yes' answer to a particular question and [2] to denote a 'no' answer for the same question.

CIRCLE THE RELEVANT BOX, E.G. [M (1) | F (2)] OR FILL IN

1. NAME:..... 2. SEX [M (1) | F (2)] 2 ☐

3. VILLAGE:..... 4. DISTRICT:.....

5. AGE

| MONTHS | | | | | | | | | | | | YEARS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |

5 ☐☐

6. DATE OF ADMISSION:.....

7. WEIGHT ON ADMISSION:.....KG

8. VACCINATION STATUS (COPY FROM M.C.H. CARD)

| NIL | ONE | TWO | NOT KNOWN |
|---|---|---|---|
| 1 | 2 | 3 | 4 |

8 ☐

9. AGE OF VACCINATION (CIRCLE TWICE IF TWO)

| 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

9 ONE ☐☐ TWO ☐☐

10. PRESENTING SYMPTOMS

| 1 RHINITIS | 2 FEVER | 3 COUGH | 4 CONJUNCTIVITIS |
|---|---|---|---|
| 5 KOPLIK SPOTS | | 6 RASH | 7 DIARRHOEA |

10 ☐☐☐☐☐☐☐

11. COMPLICATIONS PRESENT OR DEVELOPING

| 1 PNEUMONIA | 2 BRONCHITIS | 3 LARYNGO-TRACHEITIS |
|---|---|---|
| 4 STOMATITIS | 5 GASTROENTERITIS | 6 DEHYDRATION |
| 7 SKIN SEPSIS | 8 OTITIS MEDIA | 9 MALNUTRITION |
| 10 MALARIA | 11 ENCEPHALITIS | 12 OTHER |

IF OTHER, SPECIFY:.....

11 ☐☐☐☐☐☐☐

12. DATE OF DISHARGE OR DEATH:.....

13. WEIGHT ON DISCHARGE:.....KG

14. [DISCHARGED (1) | DIED (2) | ABSCONDED (3)] 14 ☐

15. DURATION OF HOSPITAL STAY (12 - 6)

| DAYS | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | ↔ |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |

15 ☐☐

16. WEIGHT CHANGE (13 - 7)

| KG: | MINUS | | | | | | | | PLUS | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |

16 ☐☐

17. IF DIED, CAUSE OF DEATH:.....

18. P.T.O. FOR REMARKS SIGNATURE:.....

Example

As an example the different categories for usual source of drinking water may be numerically coded as follows:

| | | | | |
|---|---|---|---|---|
| 1 | ☐ | piped, in house | 5 | ☐ rain water |
| 2 | ☐ | pumped, in house or yard | 6 | ☐ spring, river or lake |
| 3 | ☐ | pumped or piped, public facility | 7 | ☐ other sources: |
| 4 | ☐ | open well | Specify:_____ | |

Pre-coding

The analysis of a study is made easier and more efficient if the data are coded at the time of collection. Data may be coded either at the time of interview, in which case the record forms need to be pre-coded, or after completion of the records (post-coding). Pre-coding saves time and is generally preferable as it may prove difficult to code information in free form. However, pre-coding is only possible if you know in advance the answer to each question. That is, questions which are 'closed' or which have predictable answers such as age or number of children. It is therefore unlikely that a record form will be entirely pre-coded.

'Others'

For questions in which fixed categories of possible answers are provided, it is often useful to include a category labelled 'others' for those answers which have not been anticipated by the investigator. However, the interviewer should seek and record specific information on an unanticipated answer, as merely ticking off the box for 'others' will not be very helpful.

Principles of coding

There are two main points to consider in order to achieve maximum efficiency in the processing of numerically coded information: simplicity and standardisation. It is good practice to use a simple single digit code for each item of information. One occasion when multiple-digit codes are necessary is when exact age is recorded. For example a 75-year-old man would be coded |7|5|, whereas a 6-year-old girl would be coded |0|6|.

Standardisation

Whenever possible, numerical codes should be standardised. Standard replies like 'yes', 'no', 'don't know' and 'not known' should each have the same codes throughout the recording form, to avoid confusion and error resulting from this classification.

Coding and design of records

Recording forms can be designed in a variety of layouts. Below are shown three ways (A, B and C) of arranging the coded responses to the question 'How many living children do you have?'

| | | | | |
|---|---|---|---|---|
| A | 0 | ☐ None | 4 | ☐ 4-5 children |
| | 1 | ☐ 1 child | 5 | ☐ 6 or more children |
| | 2 | ☐ 2 children | 6 | ☐ Not applicable |
| | 3 | ☐ 3 children | 7 | ☐ Not sure |

| | | | | |
|---|---|---|---|---|
| B | 0 | None | 4 | 4-5 children |
| | 1 | 1 child | 5 | 6 or more children |
| | 2 | 2 children | 6 | Not applicable |
| | 3 | 3 children | 7 | Not sure |

C 0 None; 1 1 child; 2 2 children; 3 3 children;
4 4-5 children; 5 6 or more children; 6 Not applicable;
7 Not sure.

There are advantages and disadvantages to each of these different layouts. In example A, a designated space or box is provided for each category, so that a tick or a cross can be made for the appropriate category. This makes the coding of information neat and elegant. However, it is time consuming to draw the boxes.

Example B sets out the coded responses in the same way but without the boxes. Here the numerical code for the appropriate category should be circled. It takes less time to produce but it uses quite a lot of space.

Example C shows how economy of space can be achieved by packing as many categories of replies as possible into one line.

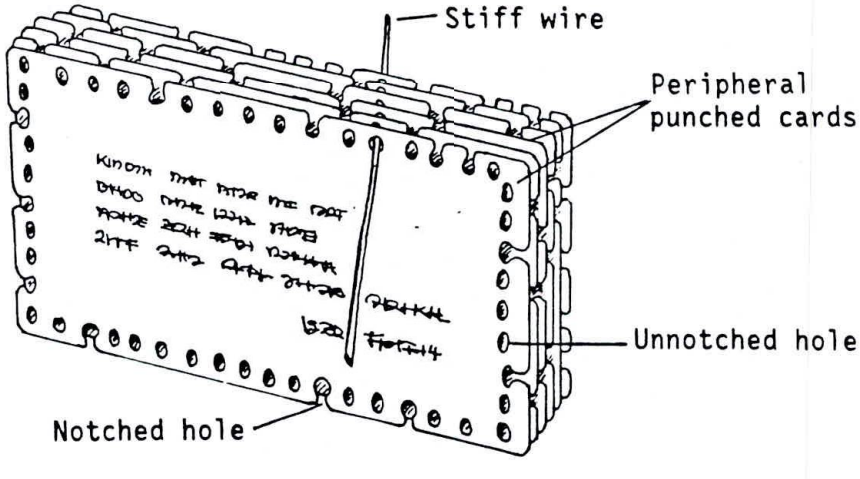
Coding column

In the design of a record form, provision is usually made for a coding column on the right hand margin of the form (see Figure 1). The reason for this is to organise the coded information relating to a respondent into a sequence in order to facilitate data processing. Take great care in the transfer of pre-coded information from the question to the appropriate box in the coding column. See Unit 5, section 5.2 for further information on how to sort coded information during data processing.

**PERIPHERAL PUNCH CARDS**

One type of record form that can greatly assist subsequent data processing is the peripheral punch card, also known as the edge-notched card. This is a specifically designed recording form with complete holes punched along its four edges (see Figure 2).

Figure 2 An example of peripheral punched cards




Uses

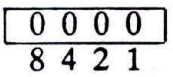
Peripheral punch cards can be used in two ways:

- As a recording form on which data and information can be entered.
- As a data-processing document through which items of information can be sorted by the notching of appropriate holes

along the edges. It is a more efficient method of data-processing than the other two manual sorting methods, hand tallying and hand sorting, which are described in Unit 5, section 5.2.

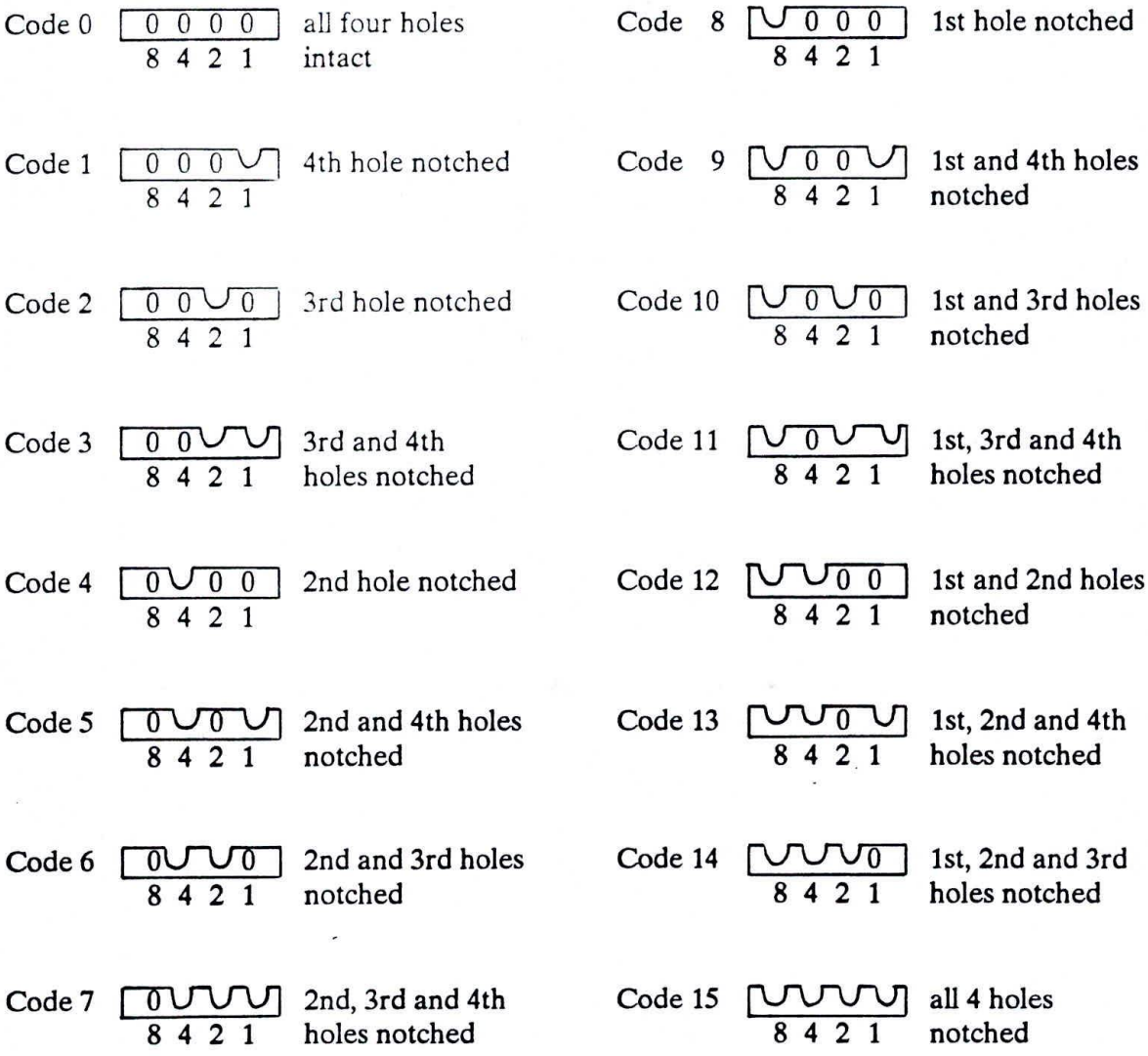
Coding

The principle used in the coding of information is simple: the holes are spaced along the edges in groups of four. In each group the holes are assigned the numerical values of 8, 4, 2 and 1 respectively, beginning from the extreme left hole.




By notching one hole or a combination of holes in each group it is possible to record all numbers from 0 to 15. Figure 3 shows how this is done. The notches can be made with a pair of scissors.

Figure 3 Coding convention for a four-holed peripheral punch card system



Sorting

Sorting of the cards is done by passing a stiff wire or knitting needle through each of the four holes in turn. When the needle is raised, the cards with intact holes remain on the spike while the others fall out. The number of cards which have dropped out represents the number of cases which have been punched for the same aspect, i.e. for the same code. Although peripheral punch cards are available commercially they can be improvised by clipping holes around the edges of a recording form with a single-hole puncher.

**IMPROVING THE RECORD FORM**

Preliminary testing

When you have designed your record form it is a good idea to discuss it with your colleagues and make modifications to it in the light of their comments. The form should then be pretested on a small number of subjects (between ten and thirty). This preliminary trial will indicate whether you need to change the questions, change the sequence or, whether you need to shorten the questionnaire. This trial may also point to a need for changes in the interviewing instructions. The test should not be carried out on members of the study population, as this may affect their responses during the actual study. However, the subjects chosen for the trial should be similar in their characteristics to members of the study population. During the trial the interviewers should record not only the information requested on the form, but also the reaction of the respondents to the questions.

Results

The results of the trial should indicate:

- questions which are offensive;
- questions which are hard to understand;
- questions which are ambiguous;
- questions which do not seem to elicit the information they are intended to obtain;
- confusing sequences of items.

If a question elicits many 'don't know' answers it is probably unsatisfactory. If it produces many qualifying comments, such as 'sometimes', 'if I have time', 'if it isn't market day' etc. the response categories are probably not suitable.

## 4.3 HOW DO I ORGANISE DATA COLLECTION?

In this section we shall look at some of the practical points which you should consider before data collection actually begins. Although some of these points may seem trivial or obvious, your study may fail if you neglect them.

### PRACTICAL PREPARATIONS

You will have to solve many practical problems which are specific to your study, the area you are working in and the people you are studying. You may find it helpful to make a checklist for yourself of the main points which you should consider when organising your data collection.

Checklist

Here are a few general points to think about and relate to your own community.

Obtaining co-operation

Good co-operation between the sruvey team and the population being surveyed is essential if there is to be the required response and coverage. People will participate if they want and support the investigation and have no unjustified fears about it. Therefore, they must be adequately informed about it, perhaps at an initial community meeting. This information must lead to motivation to participate, either from individual rewards, such as prompt treatment of illness and easy referral to hospital, or indirect rewads such as a feeling of benefiting the community.

Selection of personnel

The survey personnel may themselves arouse adverse reactions in certain cultures and societies, especially if they are foreign to the area, unable to speak the local language or are unaware of the sensitivities of the people. It is often an advantage if some of the staff are known in the area and have been locally recruited.

Laboratory tests

Where laboratory tests are to be undertaken, consider the reaction of the population. People may object to putting a specimen of faeces in a container or be afraid when blood samples are taken away for examination. Nowadays it is often easier to preserve and transport specimens to a big laboratory rather than attempt to process them under adverse conditions in the field. Containers must be cheap, durable, and have labels that do not come off or become illegible.

'Line of flow'

If a survey includes a number of procedures, each done by a different worker, it is advisable to devise a 'line of flow' whereby patients or individuals pass from one 'station' to another and have a different procedure done at each. If children are being examined, it is wise to have traumatic procedures last: blood taking, for example, should come after examination of the skin.

Equipment

Equipment for field surveys has to be carefully listed, as once out in a remote rural area it may be impossible to get an omitted

item. Each procedure must have a checklist of required items, e.g. instruments, containers, forms, reagents, labels.

**TRAINING PERSONNEL**

One aspect of organisation is the training of personnel who will actually be responsible for collecting the data. This is important for the following reasons.

- To motivate the workers to a high level of efficiency and interest. Poorly motivated staff are unlikely to provide reliable data. Also, the study population will notice very quickly if a recorder lacks interest and expertise, and their cooperation will suffer accordingly.
- To improve the technical skills of the participating personnel in order to reduce observer variation and ensure a fair degree of standardisation as we discussed in Unit 3, section 3.5.

You might use the following training methods.

Training sessions
- Hold regular classroom and field sessions for the people who will be collecting the data.

Discussion
- Encourage free discussion and comments. Many potential field problems may be identified during these sessions, especially when experienced field staff are present.

Practical test
- Give a practical test at the end of the training session to make sure that your staff have attained the required level of skill.

Equipment
- Give training on the correct use of any equipment which is to be used.

Involvement
- Familiarise the workers with the purpose, objectives and general components of the investigation. In particular, stress the importance of the study and the benefits that it would bring to the community. This will help to make the field worker feel that he or she is an essential member of an important and useful study. he workers should be be able to make the community feel involved and encourage them to cooperate throughout the study.

Questionnaires
- If data are to be collected through a questionnaire, workers should thoroughly understand the relevance of each question and how it is to be asked. When these are understood, the workers should practis using the questionnaire on each other and on their own families.
  Check how successful these interviews are and find out if there are any problems.

Manuals
- Operational manuals may be very useful for the field worker, particularly for those working in remote areas with no immediate access to supervising staff. Prepare enough manuals before the staff are sent out.

**A PILOT STUDY**

You have decided what to record and how, and from whom and when. You have obtained approval from the authorities and you have fully informed the community about what will take place and why. You have trained the field personnel, and you have obtained and tested the necessary equipment, as well as the recording form.

Purpose

At this stage it will be a good idea to do a limited pilot study under field conditions to measure the appropriateness of the plans and the feasibility of actually achieving what is being attempted. At this stage there is still time to correct mistakes, to fill gaps or to make adjustments in the way the information will be collected or measurements and laboratory procedures completed.

A pilot study can also help you to solve logistical and administrative problems and to plan a realistic time schedule.

If the results of the pilot study show that your plans need radical change, then the pilot study may have saved the main programme from failure. If no major adjustments are necessary, the information obtained in the pilot study can be incorporated into the main data base.

Size

When the objective of the pilot study is merely to demonstrate that the method of collecting data is a reasonable one, the size of the sub-sample can be very small. As few as ten subjects can show you where any major problems are likely to occur. The pilot study needs to be larger if its objective is to test the original hypotheses and thus demonstrate whether or not significant data are likely to be uncovered by a full-scale investigation.

## 4.5 SUMMARY

The main points which were covered in this unit are summarised below.

Record forms

1 A carefully designed record form will:
- Facilitate the collection of reliable data.
- Facilitate storage, sorting and analysis.

2 When designing a record form:
- Ensure that all relevant information is included.
- Avoid lengthy forms.
- Arrange questions in order of difficulty.
- Phrase questions clearly and simply.
- Ensure questions are answerable.
- Plan the layout.

3 Data analysis is made easier if the data are coded at the time of collection.

4 Data processing can be made easier by the use of peripheral punch cards.

Organisation

5 The following aspects need to be considered when organising data collection:
- Obtain cooperation of local population and medical staff.
- Select and train personnel.
- Arrange necessary laboratory facilities.
- Work out 'line of flow', and design and print record cards.
- Obtain special equipment (with spares), drugs, other medical supplies, etc.
- Organise transport and accommodation.
- Try out survey procedures to assess acceptability to the local population and to test techniques.

Monitoring

6 When data collection is in progress, you will need to:
- Supervise staff and ensure continuing accuracy of observation and recording.
- Ensure continuing cooperation of the population.

# UNIT 5: DATA ANALYSIS: SORTING AND SUMMARY

# UNIT 5: OBJECTIVES

Study with this unit will enable you to analyse data by sorting and summarising the information.

In particular you will be able to:

- Select an appropriate method for sorting raw data.
- Describe the main types of data and select appropriate methods for their analysis.
- Tabulate data in order to examine frequency distributions.
- Select an appropriate method for presenting frequency distributions diagrammatically.
- Define and calculate indices to summarise quantitative data.
- Tabulate data in order to examine the relationships between two or more variables.
- Interpret the relationship between pairs of variables using a scatter diagram.

---

# UNIT 5: CONTENTS

## 5.1 INTRODUCTION

Raw data

Data collection leads to the accumulation of a mass of unprocessed or 'raw' data. These raw data then need to be processed or sorted in order to extract the relevant information so that the study objectives can be achieved.

Data summary

The process of organising the raw data into a compact and readily comprehensible form is known as data summary. Data summary usually involves the preparation of two kinds of tables:

Tabulations

- Frequency distributions, which show the number of individuals in each of the categories by which a variable is measured (for example age in years).
- Contingency tables or cross tabulations, in which each individual is simultaneously classified according to two or more variables (for example age and sex).

Sorting

Before these tables can be prepared the raw data need to be sorted using one of the simple methods described in section 5.2.

Frequency distributions

A good way to begin summarising the data is by examining the frequency of each variable. Frequency distributions provide a way of organising a collection of measurements so that we can determine what levels are common and what levels are rare. They can be presented in the form of tables or diagrams. Diagrams can often show patterns and trends more clearly than tabulations can.

Diagrams

Type of data

As the methods you use for summarising and presenting the data will depend on the nature of the data, we will first consider the main types of data which you are likely to have. In particular, whether they are qualitative or quantitative (section 5.3). We will then help you to choose an appropriate method for examining the frequency distributions of qualitative data (section 5.4) and quantitative data (section 5.5).

Summarising indices

Section 5.6 describes simple indices which summarise the frequency distributions of quantitative data, such as the mean and standard deviation.

Cross-tabulations

Tables involving only one variable are known as simple or one-way tables. A major part of the analysis usually involves pairs or sets of variables rather than single variables (section 5.7). Cross-tabulation or contingency tables can be used to obtain the frequency distribution of one variable by subsets of another variable, such as age or sex. Cross-tabulations may also reveal associations between variables, which can then be examined in more detail using analytical methods (see Unit 6).

---

## 5.2 SORTING THE DATA

Sorting methods

There are several simple data processing techniques which do not require a computer or other sophisticated equipment. Although the simpler methods can be tedious and less satisfactory for processing large amounts of data, they do have the advantage of keeping you 'close to the data' so that you will become familiar with the information you are working with. We have already considered one method for processing data, the peripheral punch card, when we looked at the design of record cards in Unit 4, section 4.2. In this section we shall consider two very basic methods: hand tallying and hand sorting.

Peripheral punch cards

### HAND TALLYING

This is the simplest method of sorting data. A tally sheet is prepared in the form of a skeleton table, and a tally mark is made in the appropriate space or cell for each individual (Figure 1). The usual method is to make a vertical mark for each individual; every fifth mark is drawn diagonally across the preceding four: 𝍸. This indicates a group of five items and makes subsequent counting easier. Figure 1 shows a tally sheet for a two-way tabulation (age and sex).

Counting by fives

Hand tallying is a convenient method provided the amount of data is not too large. The disadvantages are:

Disadvantages

- It is laborious and time consuming.
- Errors are likely to occur because processing large amounts of data by tallying is tiring.

Errors

- Common errors in tallying include misclassification of an item of information (for example ticking off 'male' instead of 'female'), double counting an item, or missing it. The risk of making an error is even higher if a complicated cross-tabulation is used.
- If an error is detected (for example, if the total number shown in the table is one less than the actual number of individuals!) the entire tallying process has to be repeated.

Figure 1 Age and sex distribution of cholera cases admitted to four treatment centres in Kyela district.

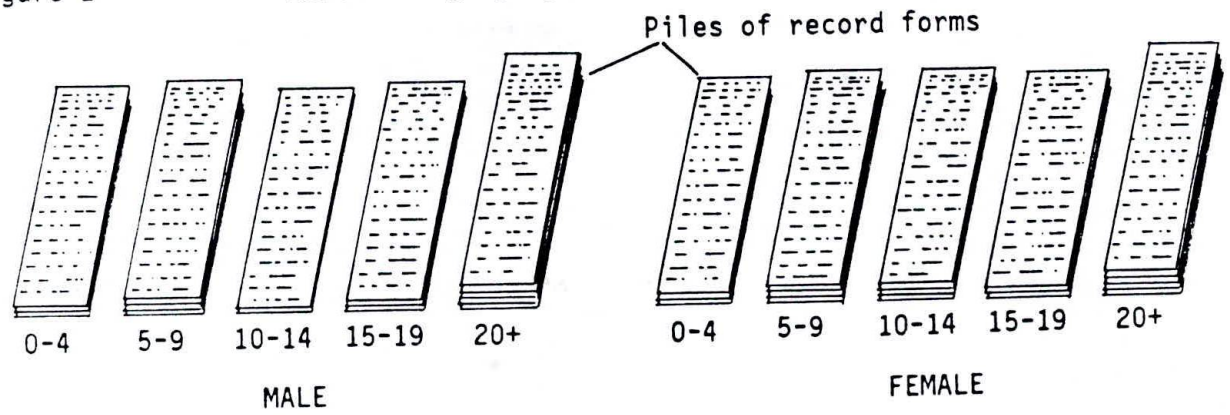
| Age | Male | | Female | | Totals | |
|---|---|---|---|---|---|---|
| 0-4 | 𝍸 𝍸 𝍸 1111 | 19 | 𝍸 𝍸 111 | | 13 | 32 |
| 5-9 | 𝍸 11 | 7 | 𝍸 𝍸 | | 10 | 17 |
| 10-14 | | 0 | 1 | | 1 | 1 |
| 15-19 | 111 | 3 | 111 | | 3 | 6 |
| 20 & over | 𝍸 𝍸 𝍸 𝍸 | | 𝍸 𝍸 𝍸 𝍸 𝍸 | | | |
| | 𝍸 𝍸 𝍸 𝍸 | | 𝍸 𝍸 𝍸 𝍸 𝍸 | | | |
| | 𝍸 𝍸 𝍸 11 | 57 | 𝍸 𝍸 𝍸 𝍸 𝍸 | | | |
| | | | 𝍸 𝍸 𝍸 𝍸 𝍸 | | | |
| | | | 𝍸 𝍸 𝍸 𝍸 𝍸 | | | |
| | | | 𝍸 𝍸 1 | | 136 | 193 |
| Totals | | 86 | | | 163 | 249 |

**HAND SORTING**

This technique is similar to hand tallying in that counts are made of specific items of information as they appear on the data sheet, and the numbers are entered into a skeleton table. Hand sorting requires a separate and easily handled record for each individual. The records are sorted and physically separated into piles conforming with the cells in the skeleton table. The numbers of records in each pile (corresponding to the number of individuals in that category) are then counted, checked and entered.

To prepare a cross-classification the records are sorted into piles according to one variable, then each pile is sorted according to a second variable, and so on until the cross-classification is complete. Figure 2 shows the sorting of records according to age and sex.

Figure 2 Handsorting by age and sex




The records you will use may be the forms or cards on which the data were initially recorded, or special forms or cards onto which the data have been transferred for this purpose.

Errors

Just as in tallying, sorting is susceptible to error when we have to hand sort data for two-way tables. However, unlike tallying, errors arising from misclassification can be checked and rectified without the need to repeat the entire procedure again. For example, if in checking through a pile of records of males aged 0-4 years you find a record belonging to a female of the same age, all you need to do is transfer the record to the correct pile.

Many people find that hand sorting is a satisfactory method except in very large or complicated studies.

## SORTING CODED INFORMATION

In Unit 4 we described how recording forms can be coded so that each item of information on the form is assigned a number. Figure 3 shows an example of part of a coded record form, and Figure 4 illustrates how the coded information on this form can be transferred onto a 'master' tally sheet. Using home-made tally sheets such as these, you can transfer all the information from each recording form onto one large sheet of paper (join several sheets together if necessary). If you transfer the information as soon as you receive the record form, you can quickly identify errors or missing data, and take steps to rectify the situation.

When you have completed the tally sheet you can easily count the numbers in each category and prepare frequency tables for each variable (see below).

Figure 3 Part of a coded record form

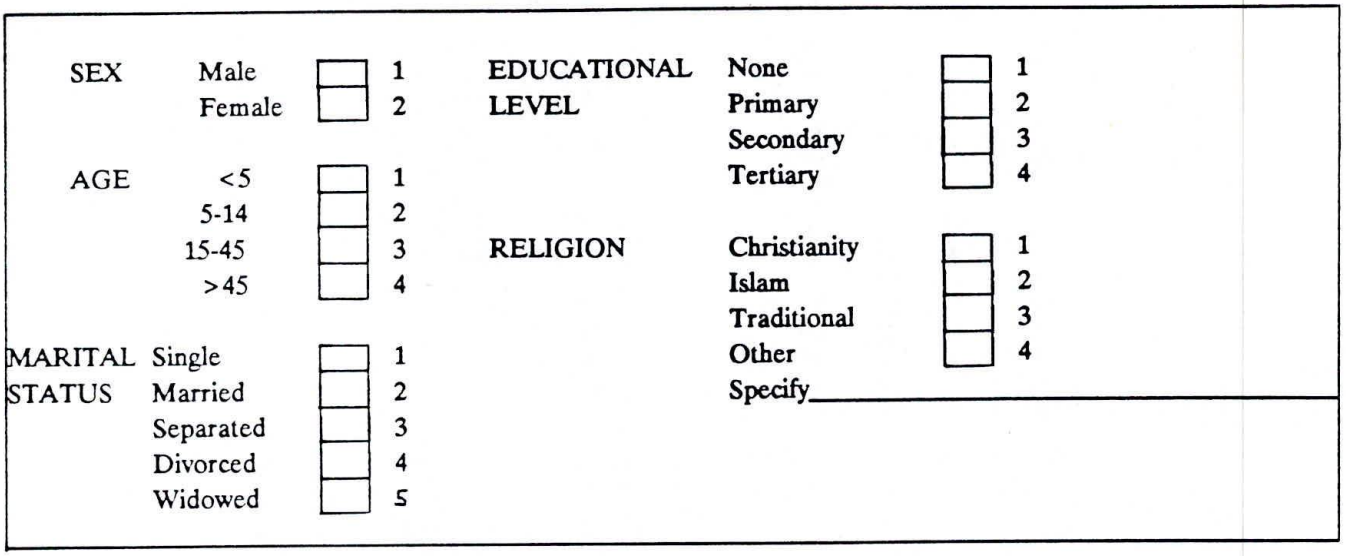


Figure 4 Tally sheet for coded information

| SERIAL NO. | SAMPLE AREA No. | HOUSE NO. | HOUSEHOLD NO. | SEX | | AGE | | | | MARITAL STATUS | | | | | EDUCATION | | | | RELIGION | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| | | | | | | | | | | | | | | | | | | | | | | |

## 5.3 TYPES OF DATA

The nature of the data you have collected determines what methods you will use for summarising and presenting the data and also the methods for statistical analysis (see Unit 7). We will therefore briefly consider the two main types of data: qualitative and quantitative.

### QUALITATIVE DATA

Qualitative (or categorical) data are provided by variables that generate information which represents counts, not measurements. An example is sex because data are derived by counting the number of individuals who are male and the number who are female. Other examples could be occupation, cultural group or source of drinking water. Qualitative data can be further subdivided into nominal and ordinal data (Figure 5).

Figure 5 Types of qualitative data.

| Type | Examples |
|---|---|
| Nominal data (Mutually exclusive unordered categories) | • Sex<br>• Blood groups<br>• Occupation<br>• Religion<br>• Ethnic group |
| Ordinal data (Related categories ranked in some order according to preference, difficulty, etc.) | • Educational level (none, primary, secondary, tertiary)<br>• Social status<br>• Lack of food for nursing mothers (critical, severe, moderate, slight)<br>• Opinion about the health service (excellent, very good, adequate, poor) |

**QUANTITATIVE DATA**

Quantitative (or numerical) data are provided by variables that generate measurement data. Examples are height, weight, pulse rate or white blood cell counts etc. because information or data from such variables are derived from actual measurements. Quantitative data can be further subdivided into discrete and continuous data (Figure 6).

Figure 6 Types of quantitative data

| Type | Examples |
|---|---|
| Discrete data (Discrete variables have a finite number of values in any given interval. They are obtained by counting and are usually whole numbers) | • Number of children in families<br>• Duration of illness<br>• Number of clinic attendances<br>• Number of beds in a a hospital ward<br>• White blood cell counts |
| Continuous data (Continuous variables have potentially an infinite number of possible values in any interval. They are usually expressed as fractions or decimals rather than whole numbers) | • Height (in metres)<br>• Weight (in kg)<br>• Blood urea levels<br>• Body temperature<br>• Skinfold thickness |

Precision

Note that continuous variables such as height, can never be measured with complete precision because of the coarseness of our measuring instruments. The measure is made with reference to divisions on a scale and there is a practical limit to how fine we can make the divisions. We might measure height to the nearest 0.5 cm or 1.0 cm. The divisions we use will depend on the sensitivity of the measuring instruments and on how precise we need the recorded measurements to be.

## 5.4 FREQUENCY DISTRIBUTIONS: QUALITATIVE DATA

### FREQUENCY TABLES

Frequency distributions can be examined by preparing a separate table for each variable, showing how many individuals fall into each category (qualitative data) or at each value (quantitative data) of the variable.

Construction

The construction of frequency tables to show the distribution of qualitative data is very easy. All that is required is a count of the number of observations for each of the possible values for the variable in question. For example, to obtain a frequency distribution for sex, you would count the numbers of males and the numbers of females in the sample. Where applicable you may also want to record the number of 'Don't knows' or queries.

Example

Figure 7 shows the frequency distribution of blood groups among a sample of 200 persons. Note that percentages (relative frequency) can be used instead of, or as well as, actual numbers. This is especially useful when large numbers of people are investigated or when you want to compare the results in two groups of different sizes.

Figure 7 Distribution of blood groups among 200 persons

| Blood group | Number | Percentage |
|---|---|---|
| 0 | 90 | 45% |
| A | 82 | 41% |
| B | 20 | 10% |
| AB | 8 | 4% |
| Total | 200 | 100% |

Decisions

When you are constructing a table, you need to decide:

- What variable or set of variables will be covered by the table.
- What categories will be used.
- How the table will be arranged.
- What individuals will be included or excluded.

Classifications

When constructing a table, two conditions must be satisfied.

- There should be no ambiguity as to which classification an item of raw data should be entered (i.e. the classifications used in a table must be mutually exclusive).
- No item of raw data should be left out of the tabulation (i.e. the table classifications must be exhaustive). For example, an enquiry may uncover only two children with a congenital condition, but if this may be at all relevant it must appear as a category.

Labelling

All tables must be clearly labelled, even if they are only rough tables for your own use. It is surprising how quickly you will forget what you have done, especially if you are only able to work on the data for short periods at a time. Make sure you mark on the table exactly what variables are tabulated and their scales of measurement.

Decisions about analysis

The information which you obtain from frequency tables may influence the decisions you make about subsequent steps in the analysis. For example you may have to abandon plans for detailed studies of relationships between illness and occupation if you find that most of the people in the community you are studying in fact have the same occupation (i.e. make sure your variables do vary). If you find only three cases of leprosy, there is no point in planning complicated tables on the epidemiology of the disease. If there are an excessive number of individuals in the 'unknown' category, you may have to decide that the variable cannot be studied. You may discover oddities in the frequency distribution which may lead you to suspect that the data are not accurate, for example an extreme maximum or minimum figure.

## FREQUENCY DIAGRAMS

As an alternative to tabulation, data on frequency distributions are often represented in a diagram because patterns of distribution are often more readily discerned and compared by using diagrams instead of tables. The type of diagram you would want to draw depends on the type of information it is supposed to present.

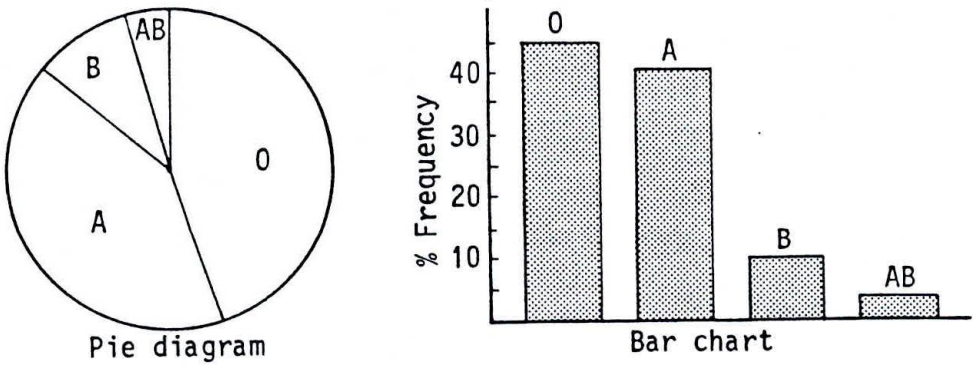
Type of diagrams

The two commonest types of diagrams used to represent the frequency distributions of qualitative data are:

- bar charts
- pie diagrams.

We shall now consider how these diagrams may be used. For information on actually how to draw bar charts and pie diagrams, refer to Unit 8 section 8.3. Figure 8 shows how the blood group data from Figure 7 can be illustrated using a pie diagram or a bar chart.

Figure 8 Frequency diagrams to illustrate qualitative data: blood groups in a population



Pie diagrams

A pie diagram consists of a circle whose area represents the total frequency and which is divided into segments representing the frequency of various groups or divisions of a descriptive attribute.

Bar charts

The main features of a bar chart are:

- The diagram consists of bars or columns.
- The width of the bar has no meaning and frequencies are represented by the length of the bars.
- The bars are of equal width and there are spaces between consecutive bars because the horizontal axis deals with information that is non-continuous in nature.

Orientation

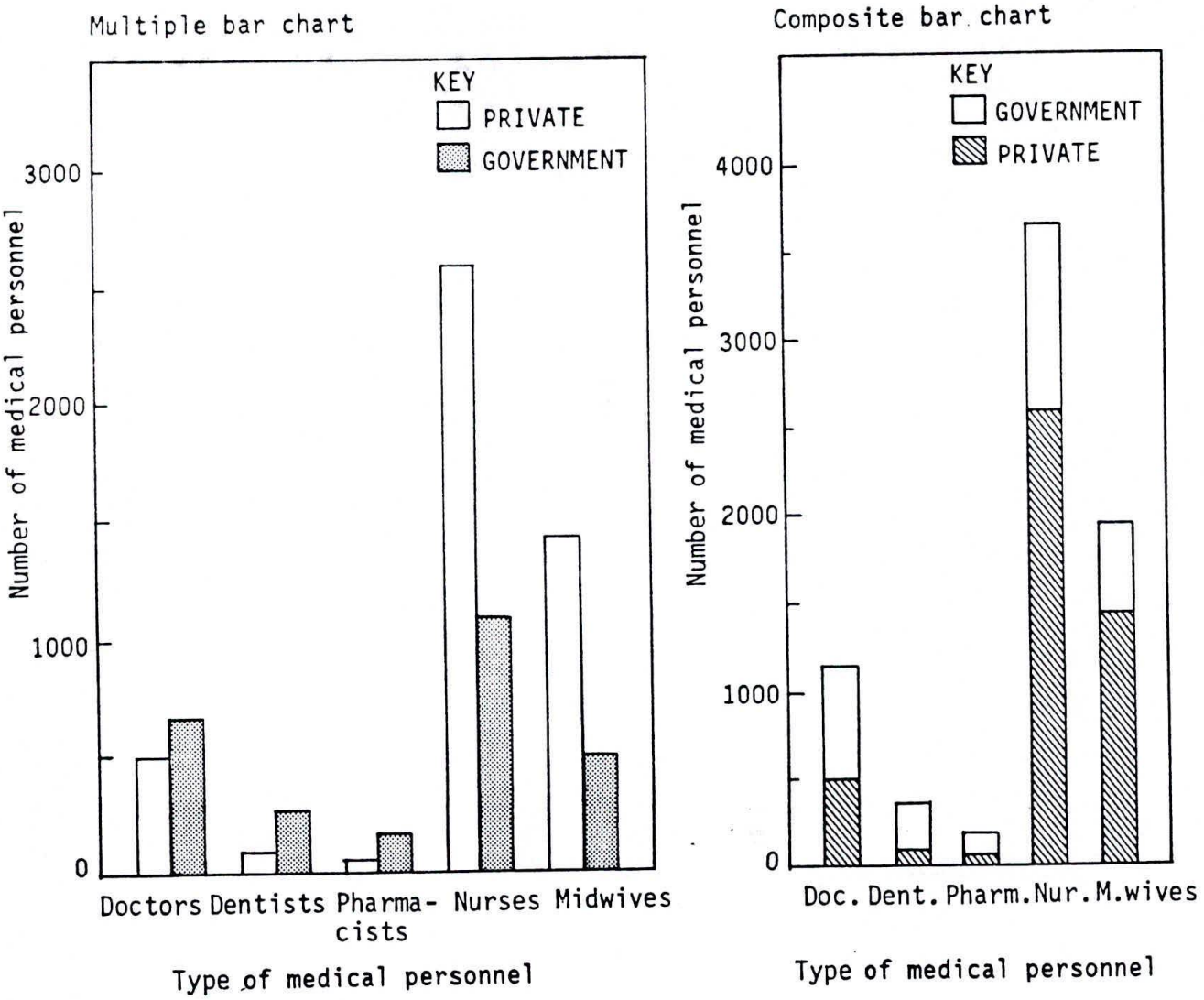
It is usual to have the variable or attribute on the horizontal axis and frequency on the vertical axis, so that the bars are vertical.
Occasionally the diagram works better if the axes are reversed so that the bars are arranged horizontally.

Comparing data

Bar charts can also be used to compare visually two or more frequency distributions as shown in Figure 9 overleaf.

- Multiple bar charts use multiple (usually two) bars for each variable.
- Composite bar charts use a single bar for each variable.

Figure 9 Bar charts to show the distribution of registered medical personnel in government and private practices in Singapore, 1968



## 5.5 FREQUENCY DISTRIBUTIONS: QUANTITATIVE DATA

### FREQUENCY TABLES

Discrete data

The examination of one-way distributions of quantitative data which are discrete is exactly the same as for qualitative data (see page 151) Figure 10 shows how to present the frequency distribution of a discrete variable - the number of children in each of 100 families.

Figure 10 Frequency distribution of 100 families according to number of children.

| No. children in family | Frequency (No. families) |
|---|---|
| 0 | 8 |
| 1 | 12 |
| 2 | 14 |
| 3 | 16 |
| 4 | 13 |
| 5 | 10 |
| 6 | 8 |
| 7 | 6 |
| 8 | 4 |
| 9 | 4 |
| 10 | 2 |
| 11 | 2 |
| 12 | 1 |
| Tota1 | 100 |

Continuous data

For continuous data the presentation is not quite so simple. If the range is relatively small, it may be practicable to list the values observed and the number of subjects with each observed value. With a wide range and many observations it will be better if the data are grouped and the numbers in each group are considered. This is the situation in Figure 11.

Figure 11 Distribution of serum uric acid levels in 267 healthy male blood donors.

| Uric acid (mg/100 ml) | Number of men | Percent of total | Cumulative percent of total |
|---|---|---|---|
| 3.0-3.4 | 2 | 0.8 | 0.8 |
| 3.5-3.9 | 15 | 5.6 | 6.4 |
| 4.0-4.4 | 33 | 12.4 | 18.7 |
| 4.5-4.9 | 40 | 15.0 | 33.7 |
| 5.0-5.4 | 54 | 20.2 | 53.9 |
| 5.5-5.9 | 47 | 17.6 | 71.5 |
| 6.0-6.4 | 38 | 14.2 | 85.8 |
| 6.5-6.9 | 16 | 6.0 | 91.8 |
| 7.0-7.4 | 15 | 5.6 | 97.4 |
| 7.5-7.9 | 3 | 1.1 | 98.5 |
| 8.0-8.4 | 1 | 0.4 | 98.9 |
| 8.5-8.9 | 3 | 1.1 | 100.0 |

Grouping data

To group data you should first identify the highest and the lowest values and then divide up the difference between them into equal intervals. You may choose any interval of grouping depending on the number of observations, but an interval which results in five to 15 groupings is often the most suitable. Less than five groupings compresses the pattern of distribution too much, whereas if you use more than 15 groupings you may have so few observations in each group that the pattern of distribution is obscured.

Example

Therefore to obtain a frequency table for continuous data as in Figure 11, we first define the groups (as shown in the first column). We then record the number in each group (as shown in the second column). Thus the first two columns give the frequency distribution of serum uric acid levels in the 267 men.

Relative frequency

We can obtain a relative frequency distribution (e.g. a percentage distribution) by dividing the number in the class by the total number and and multiplying by 100 (as shown in the third column). Notice that this column gives the distribution in a 'standard' form because it may now be compared with similar distribution data for a collection that differs in size (for example, data for 323 healthy men).

Cumulative frequency

The fourth column gives the cumulative frequency distribution. This enables us to make a quantitative statement about a given level of the measurement. For example, we can say that about 92% of the men had uric acid levels below 7 mg/100 ml.

## FREQUENCY DIAGRAMS

There are three types of diagrams which may be used to represent the frequency distributions of quantitative data:

- frequency histograms
- frequency polygons
- cumulative frequency polygons.

Figure 13 overleaf shows how these three types of diagram can be used to illustrate the data on serum uric acid levels in Figure 11.

Frequency histograms

Frequency histograms are the commonest type of diagram for presenting quantitative data which have been grouped, such as age groups or events occurring over a period of time, for example cases per day in an epidemic.

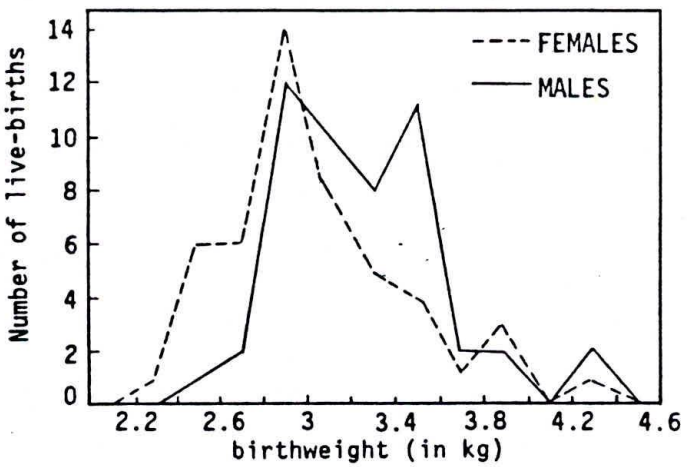
Note that the horizontal axis of the graph gives a continuous scale of the measurement variable while the vertical axis measures frequency. The area of each bar or column is directly proportional to the frequency of that class. Therefore when equal class intervals are used the frequency of each class is represented by the height of the bar, since the width is constant. The total area of all the bars corresponds to the total frequency of the distribution.

Are you sure that you understand the difference between a histogram and a bar chart? Look back at Figure 8 and the accompanying explanation if you are not.

Frequency polygons

Frequency polygons are a type of line graph in which the vertical axis represents frequency and the horizontal scale represents the method of classification. Frequency polygons are especially useful for comparing two or more distributions involving continuous measurement data, as in Figure 12, when superimposing one frequency histogram on top of another would result in a confusing diagram.

Figure 12 Frequency polygons to show the distribution of 99 live-births by sex and birthweight

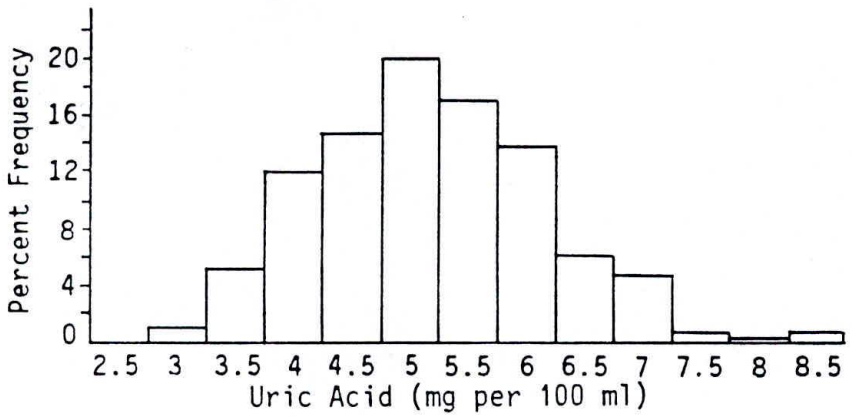



Cumulative frequency polygons

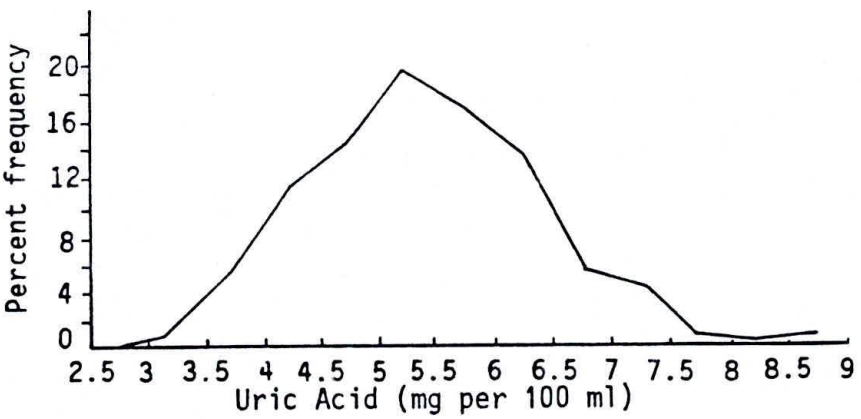
Cumulative frequency polygons (sometimes called ogives) are drawn for a distribution whose frequency has been successively cumulated. This type of diagram can be used for calculating percentile values for a given distribution.

Figure 13 Frequency diagrams to show the distribution of serum uric acid levels in 267 healthy male blood donors

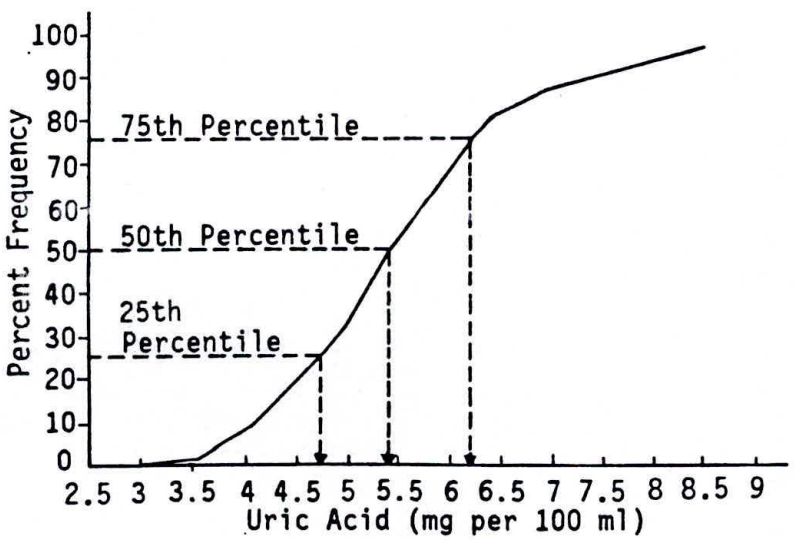
Frequency histogram




Frequency polygon




Cumulative frequency polygon

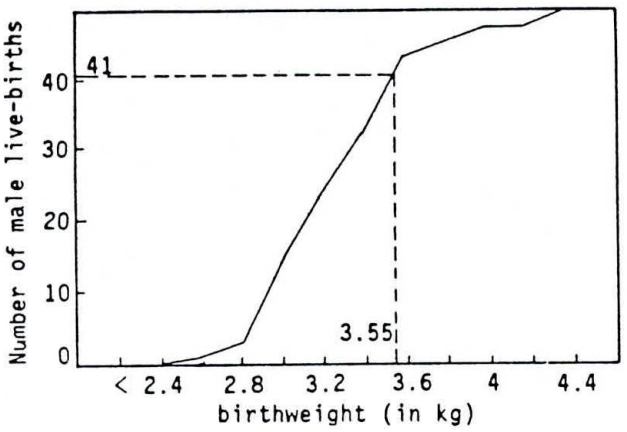


Percentiles

A percentile is the level of the measurement below which a specified proportion of the distribution falls. For example in Figure 13, 25% of the study population have a uric acid level below 4.7 mg/100 ml. 4.7 mg/100ml is thus the 25th percentile of that distribution (see also page 167).

Cumulative frequency polygons can also be used for estimating or predicting values as shown in Figure 14.

Figure 14 Cumulative frequency distribution of male live-births by birthweight




By looking at the distribution we can estimate the number of infants below a given birthweight, say 3.55 kg, by reading the value on the vertical axis (41) at the point on the cumulative frequency polygon corresponding to a horizontal axis value of 3.55 kg.

**EXERCISES**

The data in Figure 15 overleaf were obtained from a cross-sectional survey on diastolic blood pressure (DBP) in mmHg of 100 schoolchildren.

*Exercise 1* *Construct a table to show the age distribution of the sample.*

Figure 15 Diastolic blood pressure, age and sex of 100 schoolchildren

| DBP | Age | Sex | DBP | Age | Sex | DBP | Age | Sex |
|---|---|---|---|---|---|---|---|---|
| 83 | 18 | M | 74 | 18 | M | 69 | 18 | M |
| 68 | 17 | F | 73 | 18 | F | 77 | 17 | F |
| 70 | 18 | F | 70 | 18 | F | 67 | 18 | F |
| 63 | 18 | F | 68 | 18 | F | 68 | 18 | F |
| 65 | 17 | M | 68 | 17 | M | 74 | 17 | M |
| 75 | 17 | M | 75 | 17 | M | 62 | 18 | M |
| 57 | 17 | M | 56 | 17 | M | 50 | 18 | M |
| 88 | 18 | M | 72 | 18 | M | 66 | 17 | M |
| 62 | 17 | M | 61 | 16 | M | 75 | 17 | M |
| 71 | 18 | M | 70 | 17 | M | 54 | 18 | M |
| 74 | 16 | M | 75 | 18 | M | 75 | 17 | M |
| 58 | 18 | M | 49 | 17 | M | 77 | 18 | M |
| 68 | 13 | F | 52 | 12 | F | 70 | 13 | F |
| 72 | 14 | F | 72 | 12 | F | 79 | 13 | F |
| 83 | 13 | F | 56 | 14 | F | 59 | 13 | F |
| 68 | 14 | F | 67 | 14 | F | 56 | 14 | F |
| 66 | 13 | F | 61 | 13 | F | 69 | 13 | F |
| 72 | 14 | F | 70 | 14 | F | 61 | 14 | F |
| 68 | 13 | F | 77 | 14 | F | 73 | 15 | F |
| 68 | 13 | F | 62 | 13 | F | 62 | 12 | F |
| 58 | 13 | F | 68 | 14 | F | 66 | 14 | F |
| 54 | 14 | F | 77 | 16 | F | 69 | 15 | F |
| 58 | 13 | F | 55 | 13 | F | 58 | 16 | F |
| 56 | 15 | F | 61 | 16 | F | 69 | 15 | F |
| 69 | 16 | F | 72 | 15 | F | 66 | 15 | F |
| 64 | 16 | M | 61 | 15 | M | 62 | 14 | M |
| 71 | 14 | M | 50 | 14 | M | 66 | 16 | M |
| 56 | 14 | M | 60 | 15 | M | 70 | 14 | M |
| 65 | 16 | M | 64 | 15 | M | 64 | 17 | M |
| 63 | 14 | M | 72 | 17 | M | 60 | 14 | M |
| 64 | 15 | M | 72 | 14 | M | 74 | 16 | M |
| 60 | 14 | M | 67 | 16 | M | 67 | 16 | M |
| 64 | 17 | M | 68 | 16 | M | 61 | 15 | M |
| 61 | 16 | M | | | | | | |

*Exercise 2* *Construct a table to show the distribution of diastolic blood pressure readings.*

Figure 16 shows measurements of skinfold thickness in millimetres at the triceps mid-point for 121 male subjects.

Figure 16 Skinfold thickness (mm)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 11.4 | 15.3 | 9.1 | 18.4 | 10.9 | 4.7 | 9.6 | 20.6 | 10.4 | 20.5 | 22.4 |
| 14.3 | 11.7 | 11.4 | 12.7 | 18.2 | 15.1 | 14.6 | 25.3 | 11.5 | 13.2 | 7.9 |
| 12.6 | 13.9 | 16.8 | 11.4 | 27.3 | 16.3 | 13.9 | 13.2 | 11.9 | 20.0 | 13.2 |
| 9.4 | 18.9 | 10.7 | 14.8 | 17.8 | 10.8 | 16.0 | 15.7 | 17.7 | 13.5 | 11.5 |
| 11.1 | 9.6 | 15.1 | 13.6 | 13.6 | 8.6 | 6.9 | 19.1 | 18.7 | 10.1 | 16.0 |
| 20.4 | 7.9 | 16.6 | 18.5 | 16.2 | 17.4 | 18.8 | 12.6 | 22.0 | 9.6 | 11.1 |
| 15.7 | 23.7 | 13.3 | 4.9 | 8.3 | 20.1 | 15.5 | 23.1 | 10.2 | 10.7 | 15.8 |
| 17.6 | 21.3 | 16.2 | 14.9 | 9.9 | 9.1 | 9.9 | 9.8 | 8.6 | 11.8 | 9.3 |
| 14.8 | 17.3 | 9.5 | 13.6 | 12.4 | 9.5 | 14.3 | 25.7 | 12.9 | 22.7 | 12.1 |
| 10.7 | 16.8 | 11.3 | 11.3 | 11.4 | 5.9 | 10.7 | 14.6 | 19.8 | 25.5 | 7.7 |
| 18.4 | 7.9 | 7.6 | 23.3 | 9.6 | 8.4 | 10.4 | 8.1 | 12.5 | 9.0 | 30.1 |

*Exercise 3*

*Using 2 mm grouping intervals construct a table showing the frequency distribution and relative frequency distribution of the skinfold thicknesses.*

Answer 1

Figure 17 Age distribution of 100 schoolchildren

| Age (in years) | No. of schoolchildren |
|---|---|
| 12 | 3 |
| 13 | 14 |
| 14 | 21 |
| 15 | 11 |
| 16 | 14 |
| 17 | 18 |
| 18 | 19 |
| Total | 100 |

In this tabulation the variable, age, is treated as a discrete measurement variable; grouping of the data is not necessary. To make the table you should first find the two extreme values from the set of raw data, i.e. ages 12 and 18, and then obtain the classes for the table.

Answer 2

Figure 18 Distribution of diastolic blood pressure readings from a sample of 100 schoolchildren

| Diastolic blood pressure (in mmHg) | No. of schoolchildren |
|---|---|
| 45 - 49 | 1 |
| 50 - 54 | 5 |
| 55 - 59 | 12 |
| 60 - 64 | 22 |
| 65 - 69 | 26 |
| 70 - 74 | 21 |
| 75 - 79 | 10 |
| 80 - 84 | 2 |
| 85 to below 90 | 1 |
| Total | 100 |

For a continuous measurement variable like diastolic blood pressure you need to group the data in order to tabulate the distribution. To construct the classes, first identify the highest and the lowest blood pressure readings. The difference between them gives the range: in this case 88-49 = 39 mmHg. As a rough guide it is a good idea to aim for about ten classes or groups so as to give the distribution a good spread. If we divide the range by 10 we get 3.9. The nearest convenient value to this is 5, and we have used this as the class interval in the table.

Answer 3

Figure 19 Frequency and relative frequency distribution of skinfold thickness

| Skinfold thickness | Frequency | % | Skinfold thickness | Frequency | % |
|---|---|---|---|---|---|
| 4- | 3 | 2.5 | 18- | 9 | 7.4 |
| 6- | 6 | 5.0 | 20- | 6 | 5.0 |
| 8- | 19 | 15.7 | 22- | 6 | 5.0 |
| 10- | 23 | 19.0 | 24- | 3 | 2.5 |
| 12- | 17 | 14.0 | 26- | 1 | 0.8 |
| 14- | 14 | 11.6 | 28- | 0 | 0.0 |
| 16- | 13 | 10.7 | 30-31.9 | 1 | 0.8 |

## 5.6 SUMMARISING QUANTITATIVE DATA

How can you compare the measurements of skinfold thickness in Figure 16 with similar measurements made in other studies? We could compare relative frequencies, but this is difficult if there are many groupings or sets of data to compare. We need to be able to summarise the main features of the frequency distribution in a few indices.

For quantitative data, summarising indices include the mean, median and mode, which provide an indication of the centre of the frequency distribution, and the range and standard deviation, which give an indication of the amount of variation present in the measurements.

### SUMMARISING INDICES

Let us work through a simple example to see what each of the summarising indices describes and how they are calculated.

Example

When a count was made of the duration in days of episodes of diarrhoea in eleven patients, the following results were obtained.
Duration in days: 3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10

Using these simple data we will define and calculate the mean, mode, median, range and standard deviation.

Mean

The mean or average value is the most commonly used measure of the central point of the data. The mean is the average of the values and is obtained by adding all the values and dividing the total by the number of values. In this example the mean is 6. The mean we have calculated here is called the arithmetic mean. In some cases use is made of the weighted mean (see page 166).

Mode

The mode is the most frequently occurring value. In this example it is 6.

Median

The median is the middle of a series of values arranged in order of magnitude, i.e. the 50th percentile. In this example the median is the 6th value, 6, since half the values exceed it and half are below it. If there are an even number of values in a series, the median is conventionally taken as the average of the middle two values.

Range

The range is the difference between the largest and the smallest value. In this example it is from 3 to 10 days.

Standard deviation

The standard deviation describes the dispersion, spread or variation within data for each sample. Mathematically the standard deviation takes into account the distance of each value from the mean value of the group. It can be calculated in the following way.

Calculation

- Calculate the difference between the value of each observation and the mean, and then square each difference. In this example it would be:

| | |
|---|---|
| Values: | 3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10 |
| Differences from mean: | 3, 2, 2, 1, 0, 0, 0, 1, 1, 2, 4 |
| Squares of differences: | 9, 4, 4, 1, 0, 0, 0, 1, 1, 4, 16 = 40 |

- Add the sums of squares and divide by the number of observations minus one (n - 1). In this example the result (= mean squared variation) is $\frac{40}{10} = 4$

- Calculate the standard deviation by finding the square root of the mean squared variation. In this example the standard deviation (SD) $= \sqrt{4} = 2$. Therefore, the standard deviation of (days of diarrhoea) from the mean is 2. As the mean is 6, then the standard deviation can be expressed as $6 \pm 2$ or from 4 to 8 days.

Interpretation

A large standard deviation shows that there is a wide scatter of observations around the mean value, while a small standard deviation shows that the observations are concentrated around the mean with little variation between observations.

A quicker calculation

In this example the squaring of each variation from the mean is very simple, but you can imagine that doing this for large numbers of observations is very laborious. Fortunately, there is a mathematical trick to make the calculation easier. Mathematicians have proved that the sum of the squares is the same as the sum of all the squared values minus the average of all the values squared. This may make more sense to you when it is represented by a formula:

$\Sigma$ (sigma) stands for 'the sum of'
If x = each value $\Sigma x^2$ = square each value and add them all together
$(\Sigma x)^2$ = add each value together and square the total
n = the number of observations

$$\text{Standard deviation} = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$$

If we apply this formula to the example we have just been using, you will find that it is not as difficult as it first seems. The values x are:
3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10 (n = 11)

If we square each value we get:
9, 16, 16, 25, 36, 36, 36, 49, 49, 64, 100

The sum of all the squares ($\Sigma x^2$) is 436
The sum of all the values ($\Sigma x$) is 66

Therefore $\frac{(\Sigma x)^2}{n} = \frac{4356}{11} = 396$

Standard deviation = $\sqrt{\frac{436 - 396}{10}} = \sqrt{4} = 2$

*Exercise 4*

*Calculate the mean and standard deviation of the skinfold thicknesses shown in Figure 16. To assist you we have calculated for you the sum of all the values, $x$ = 1705.5, and the sum of all the squares $\Sigma x^2$ = 27,073.47. Please write out in full all your calculations. That will help you to find out where you went wrong if you did not arrive at the correct answer (see next page).*

Weighted mean

On page 164 we described how to calculate the (arithmetic) mean. But consider the following example.
In counting several fields of a blood slide one observer made six counts and the other made two. Their results were:
Observer A 10, 19, 15, 18, 12, 16
Observer B 8, 12
How would you calculate the mean in this case?
Observer B made fewer counts than Observer A, so to add the mean of A (90/6 = 15) and the mean of B (20/2 = 10) and divide by two will not reflect the larger number of counts made by A. To overcome this problem we use the weighted mean, so that more weight is given to A than B.
This is calculated as follows:

$$\frac{(6 \times 15) + (2 \times 10)}{6 + 2} = 13.75$$

Answer 4

$$\text{mean} = \frac{\Sigma x}{n} = \frac{1705.5}{121} = 14.10 \text{ mm}$$

$$\text{standard deviation (SD)} = \sqrt{\frac{x^2 - (x)^2/n}{n-1}}$$

$$= \sqrt{\frac{27{,}073.47 - (1705.5)^2/121}{120}} = \sqrt{\frac{3034.38}{120}} = \sqrt{25.286} = 5.03$$

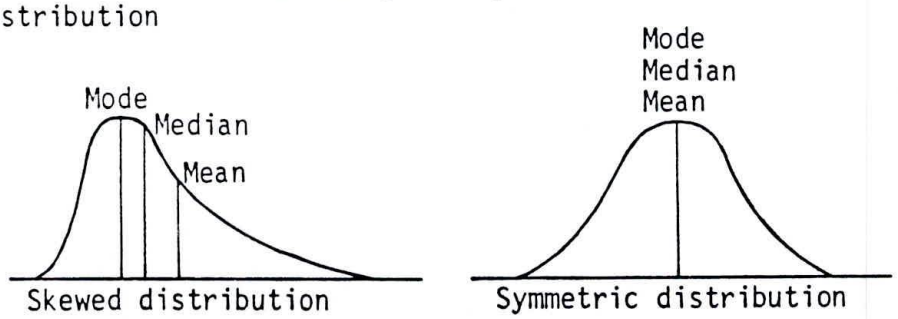
## USE OF SUMMARISING INDICES

Mean and standard deviation

The mean and standard deviation are the two usual indices used to summarise a series of measurements. However, they cannot be used automatically for all measurements because they are unsuitable if the frequency distribution is assymetrical. For any symmetrical distribution (also known as a normal distribution) the mean, median and mode will be identical. With a skewed distribution, however, these indices will be arranged as shown in Figure 20.

Note that the data on diastolic blood pressure in Figure 18 show a symmetrical distribution. The data on number of children in a family in Figure 10 show a skewed distribution.

Figure 20 Indices of central tendency for symmetric and skewed distribution




Median

In a skewed distribution the mean is an inappropriate index as it is distorted by the extreme (very high or very low) values. The median is generally preferred as it is unaffected by these extreme values. For a symmetrical distribution the mean is preferred as it uses the most information.

Mode

The mode indicates the peak of the distribution. Distributions which have two peaks (bimodal distributions) cannot be summarised using the mean and standard deviation.

Range

The range is often less valuable than the standard deviation as it only tells us about two members of a group. An extremely high or low value may be due to a measurement error. The range is the measure of variation often quoted with a median to describe skewed distributions (when the standard deviation cannot be used).

Percentiles

The median divides a group of observations into two halves with 50% of the results below the median. The median is therefore the 50th percentile (see page 164). Percentiles can be used to rank individuals in relation to the total population. The commonest

medical use for percentiles is the assessment of a child's weight for sex and age.

Normal limits

Variation is present in all biomedical measurements upon which decisions on individual patient care or community health programmes are based. We therefore need to establish standards on which decisions can be based. These standards are often referred to as 'normal values' and are generally based on measurements made on population groups classed as 'healthy'. By convention 'normal limits' are taken as 2.5th and 97.5th percentiles of the distribution of the measurement for a healthy population. If the distribution is symmetrical, these limits are equivalent to the mean $\pm$ 2 standard deviations. If the distribution is severely skewed then the 2.5th and 97.5th percentiles can be determined using a cumulative frequency plot (see page 159).

## 5.7 RELATIONSHIPS BETWEEN VARIABLES

Two or more variables

So far in this unit we have considered only single variables and their frequency distributions. However, you will probably want also to compare pairs or sets of variables. For example, you may want to look at the distribution of a variable such as diastolic blood pressure or the presence of a disease separately in the two sexes and in different age groups. In most community medicine studies the relationship between variables and these 'universal' characteristics is considered. You will also need to make those comparisons which are specified or implied in the stated objectives of the study.

Associations

Information gained from making comparisons can often suggest associations between variables which can then be examined in more detail by doing analytical studies designed to test specific hypotheses (see Unit 6).

The major method which is used for looking at relationships between two or more variables is cross-tabulation, also known as a contingency table.

### CONTINGENCY TABLES

2 x 2 tables

Contingency tables or cross-tabulations involve the use of at least two variables. In its simplest form a contingency table consists of two rows (horizontal) and two columns (vertical), excluding the row and column totals. It is therefore known as a 2 x 2 table and an example is shown in Figure 21.

The 2 x 2 table is widely used in epidemiology as a concise way of presenting data showing the relationship between variables and as the basis for statistical analysis using the chi-square test (see Unit 7, section 7.4).

Figure 21 Effect of vaccine X on the incidence of measles in community Y

| | No. of people with measles | No. of people without measles | Total |
|---|---|---|---|
| No. of people vaccinated | 3 | 27 | 30 |
| No. of people not vaccinated | 13 | 8 | 21 |
| Total | 16 | 35 | 51 |

Tables of this kind are used for qualitative data, when only the presence or absence of a characteristic is recorded, with no statement of magnitude.

Grouped data

Contingency tables can also be used for grouped data such as age. In this case the division will not be into a 2 x 2 table but a 2 x *n* table (where ***n*** is the number of groups used). See Figure 22.

Figure 22 *Onchocerca volvulus* mf carrier rate by age

| | Age in years 0-9 | 10-19 | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | Total |
|---|---|---|---|---|---|---|---|---|
| Male | 9 | 8 | 27 | 63 | 41 | 31 | 14 | 193 |
| Female | 0 | 4 | 20 | 31 | 13 | 4 | 2 | 74 |
| Total | 9 | 12 | 47 | 94 | 54 | 35 | 16 | 267 |

Percentages

The data in contingency tables can be presented as percentages of the row or column totals instead of absolute numbers. This is especially useful if, for example, the total number of diseased and control subjects are different.

Interpretation

Careful examination of the figures in the tables may reveal evidence of important differences between groups. For example, in Figure 21 it is obvious that measles occurred much less frequently in those that were vaccinated. Often this simple 'eye test' is sufficient, but if you wish to examine differences more intensively you will need to do a statistical significance test (see Unit 7).

**MULTIPLE CONTINGENCY TABLES**

Multiple contingency tables can be used to classify three or more variables simultaneously as in Figure 23.

Figure 23 Distribution of *Loa loa* mf according to age and sex

| Sex | Age | No. Examined | *L. loa* carriers |
|---|---|---|---|
| Males | 9-25 | 14 | 4 (28.6%) |
| | 26-40 | 12 | 2 (16.7%) |
| | 41-55 | 36 | 13 (36.1%) |
| | >55 | 62 | 26 (41.9%) |
| | Total | 124 | 45 (36.2%) |
| Females | 9-25 | 39 | 8 (20.5%) |
| | 26-40 | 82 | 11 (13.4%) |
| | 41-55 | 85 | 17 (20.0%) |
| | >55 | 81 | 22 (27.2%) |
| | Total | 287 | 58 (20.2%) |
| | Grand total | 411 | 103 (25.1%) |

In this case we can compare the relationship between infection rates and sex in each individual age group, or we can compare infection rates in males and females irrespective of age. Thus, we can hold one variable 'constant' while examining the relationship between others. It is often helpful to calculate separate summarising indices (e.g. means) for each stratum.

'At risk' groups

The distribution of a variable such as haemoglobin counts can vary in different groups in the population. Multiple contingency tables can help you to identify vulnerable 'high risk' population groups. If, for example, you find that anaemia is particularly frequent in women of child-bearing age, you can target health care for this section of the population.

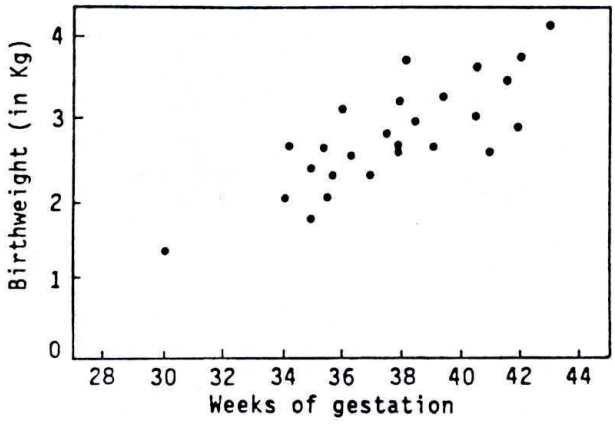
**SCATTER DIAGRAMS**

If two variables are continuously distributed, comparisons can be made more easily if the data are presented graphically rather than in a contingency table. A scatter diagram can be used when paired measurements are made of two variables. For example, a study on infant live births may provide information on the birthweight and gestational age for each infant. This information can be plotted on a graph (Figure 24) with birthweight on the vertical axis and gestational age on the horizontal axis. The resulting diagram shows

Example

the scatter or spread of the individuals in the sample with respect to the two variables. The pattern made by the scattered points indicates whether there is an association or not between the pairs of variables.

Figure 24 Scatter diagram to show the distribution of a sample of live-births by birthweight and gestational age




In Figure 24 there is a suggestion of an association between birthweight and gestational age. An infant with a high gestational age tends to be heavier than an infant with a low gestational age. When an increase in one variable is associated with an increase in the other, the two variables are said to be correlated. Statistical tests can be used to measure and test the degree of this correlation (see Unit 7). If the pattern is a random scatter of points, then probably little or no relationship exists between the two variables. The closer the points approximate to a straight line, the greater the association, whereas the direction of the line determines whether there is a positive or negative correlation.

Study advice Refer to Unit 7, section 7.6 and Reference 4.

## EXERCISES

*Exercise 5* *Using the data on diastolic blood pressure in children in Figure 15, construct a contingency table to show the relationship between blood pressure and sex.*

*What do you conclude from looking at the table?*

.....

.....

.....

*Exercise 6*

***Anaemia was found in 600 of 2000 pregnant women and in 160 of 8000 non-pregnant women. Examine the relationship between anaemia and pregnancy using a contingency table.***

***What do you conclude from the results?***

.....

.....

.....

Answer 5

Figure 25 Distribution of the 100 schoolchildren by diastolic blood pressure and sex.

| Diastolic blood pressure (in mmHg) | Sex | | Total (both sexes) |
|---|---|---|---|
| | Males | Females | |
| 45 - 49 | 0 | 1 | 1 |
| 50 - 54 | 3 | 2 | 5 |
| 55 - 59 | 4 | 8 | 12 |
| 60 - 64 | 16 | 6 | 22 |
| 65 - 69 | 9 | 17 | 26 |
| 70 - 74 | 11 | 10 | 21 |
| 75 - 79 | 6 | 4 | 10 |
| 80 - 84 | 1 | 1 | 2 |
| 85 to below 90 | 1 | 0 | 1 |
| Total | 52 | 48 | 100 |

Just by looking at this table we can see that the distribution of blood pressure values has a similar pattern in both sexes.

Answer 6

Your table should look similar to this:

Figure 26 Contingency table showing relationship between pregnancy and anaemia among 10,000 women

| | Anaemia | No anaemia | Total |
|---|---|---|---|
| Pregnant | 600 (30%) | 1400 (70%) | 2000 (100%) |
| Not Pregnant | 160 (2%) | 7840 (98%) | 8000 (100%) |
| Total | 760 | 9240 | |

Note that in addition to the actual figures we have expressed the results as percentages, which makes interpretation of the table easier.

Visual inspection of the table shows that there is a marked difference between the occurrence of anaemia in pregnant and non-pregnant women: a difference of 28% or a ratio of 15. This indicates a strong and important relationship between anaemia and pregnancy.

## ACKNOWLEDGEMENTS

The Figures used in this unit were adapted from the following sources.

- *Epidemiology for the health officer: a field manual for the tropics.* Edited by W.O. Phoon. WHO, 1985. (Figures 9, 13, 14, 15, 21 and 24).

- *A study guide to epidemiology and biostatistics.* R.F. Morton and J.R. Hebel. Aspen Publishers, Inc., 1984. (Figures 11 and 12).

- Zein, A.Z. The epidemiology of onchocerciasis in northwestern Ethiopia. *Tropical and Geographical Medicine* (1986) **38**, 33-37 (Figure 22).

- Van Hoegaerden, M. *et al.* Filariasis due to *Loa loa* and *Mansonella perstans*: distribution in the region of Okondja, Haut-Ogooue Province, Gabon, with parasitological and serological follow-up over one year. *Transactions of the Royal Society of Tropical Medicine and Hygiene* (1987) **81**, 441-446. (Figure 23).

# UNIT 6: DATA ANALYSIS: ASSOCIATION AND CAUSATION

# UNIT 6: OBJECTIVES

Study with this unit will enable you to analyse data for causal associations.

In particular, you will be able to:

- Describe the analysis of case-control and cohort studies.
- Define artifactual, non-causal and causal associations.
- Identify different types of cause and effect relationships.
- Distinguish between association and causation and list the criteria for inferring causal association.

---

# UNIT 6: CONTENTS

## 6.1 INTRODUCTION

As doctors we are concerned with limiting or preventing disease. We look for the aetiology or cause of a disease or health problem in the hope that once the cause is found, prevention will follow.

Cause

A cause of the frequency and distribution of a disease or health problem in a population is defined as a factor or habit whose reduction (or removal) leads to reduction in incidence of the disease or health problem.

Why study cause

This leads us on to why we want to identify cause even though, as you will see shortly, it is difficult to do. It is because we want to identify factors that if altered would lead to a reduction in something we consider undesirable (premature deaths, disease, etc.) or an improvement in something we regard as desirable such as health care.

Association and causation

The first step in the search for cause is the investigation of associations. An association may suggest a cause. This association may lead us to identify a component which gives an even closer association. This in turn may lead closer and closer to the ultimate cause. For example, people first noticed that malaria was associated with the bad air of swamps. This was followed by the realisation that where there were swamps there were also mosquitoes. The association between mosquitoes and malaria remained for a long time until, in the last century, the actual infecting parasite was identified in the mosquito. A series of indirect causes had thus finally led to the parasite cause.

Example of malaria

Prevention

It is often not necessary to wait until we can identify the ultimate cause before beginning preventive action. It is often sufficient just to identify an association between exposure and disease. For example, cigarette smoke has been identified as the contaminated substance that is associated with increased rates of lung and other cancers, as well as heart and respiratory disease. We do not need to identify precisely which component in the smoke is the main offender before beginning our health education programme to reduce cigarette smoking.

Investigation of cause

How, then, can we investigate the cause of a disease? The sequence of investigation follows three stages.

Observation

- Observation. A review of clinical records or the results of a descriptive health survey (see Unit 3) may show a possible association between an exposure and a disease (see Unit 5). For example, your records may show that many cases of gastrointestinal infection are occurring in a particular village. These observations may lead you to formulate hypotheses (see Unit 3) about the aetiology of the disease. For example, you may suspect that the infection is associated with a contaminated water supply, or with the way food is prepared or cooked.

Testing hypotheses

- Testing hypotheses. You can test your hypotheses relating to aetiology by doing an analytical study (see Unit 3). You would first do a case-control or retrospective study which will enable you to test the various hypotheses you have, to reject those which are unlikely and to formulate more precise hypotheses. These can then be tested by a cohort or prospective study. Cause may be due to many factors. Not everyone who swallows *Vibrio cholerae* will develop cholera.

Case-control studies

Cohort studies

Risk factors

Disease is therefore related to risk (see Unit 2). Case-control and cohort studies are designed to look for risk factors which identify those individuals who are more susceptible to disease. From this we proceed to test whether these risk factors are also causes, capable of explaining why some individuals get sick, while others remain healthy.

Experiments

- Experimentation. In some cases it is possible to demonstrate causal relationships further by an experiment or intervention study, preferably a controlled trial. Such studies seek to find out if removal or a modification of the risk factors or exposure is in fact followed by a reduction in the amount of disease.

This unit

In this unit we will focus our attention on the analysis (section 6.2) and interpretation (section 6.3) of case-control and cohort studies. These two study designs are defined and discussed in relation to planning an epidemiological study in Unit 3, section 3.3. In section 6.4 we will look more closely at causal relationships and how they can be established. Note that we will not be dealing with experimental or intervention studies.

References

We have included with this unit two references to which you should refer as you work through the sections on analysis and interpretation. Reference 1 is a retrospective case-control study to identify risk factors for the development of symptomatic cholera (by L.W. Riley *et al.*). Reference 2 is a prospective cohort study to investigate the influence of parental smoking and respiratory symptoms on the incidence of pneumonia and bronchitis in their children (by J.R.T. Colley *et al.*).

---

## 6.2 ANALYSIS OF CASE-CONTROL AND COHORT STUDIES

### GENERAL PRINCIPLES

In general, the outcome of an analytical study is the conclusion that a disease and its suspected cause are, or are not, associated. The simplest method of presenting an association is by means of a 2 x 2 (contingency) table (see Unit 5, section 5.7). Such tables are concise and easy to analyse statistically (see Unit 7), but are only appropriate if the disease and suspected cause can be recorded as present or absent (i.e. qualitative data). If the data are quantitative, for example haemoglobin counts, they must be grouped before they can be tabulated (see Unit 5, section 5.4). Associations involving continuous data which cannot be grouped must be examined using statistical indices such as the correlation coefficient (see Unit 7).

Contingency table

Correlation

In principle, both case-control (retrospective) and cohort (prospective) studies are designed to obtain the data to complete a 2 x 2 table as in Figure 1.

Figure 1 Contingency table for analytical studies

| | | Disease | |
|---|---|---|---|
| | | Present | Absent |
| Exposure | Present | a | b |
| | Absent | c | d |

Let us see how this relates to the design of case-control and cohort studies.

### CASE-CONTROL STUDIES

In retrospective or case-control studies we begin with the cases (a + c) and select a comparison group (b + d) to give the following position at the start of the study.

Study design

Figure 2 Contingency table at beginning of retrospective study

| | | Disease | |
|---|---|---|---|
| | | Present | Absent |
| Exposure | Present | ? | ? |
| | Absent | ? | ? |
| | Total | a+c | b+d |

The participants are then retrospectively assigned to the exposure rows so the 2 x 2 table in Figure 1 is then completed.

Analysis

The analysis is performed by comparing exposure rates between the case and control groups:

Exposure rate among cases $= \frac{a}{a+c}$

Exposure rate among controls $= \frac{b}{b+d}$

Example

Let us look at an example.

Suppose we are testing the hypothesis that smoking is associated with lung cancer. We take 100 cases of lung cancer and then choose 100 matched controls free of lung cancer from the general population. At the beginning of the study the 2 x 2 table is completed only to the extent shown below.

Figure 3 Contingency table at beginning of retrospective study of smoking and lung cancer

| | Cases | Controls |
|---|---|---|
| Smokers | ? | ? |
| Non-smokers | ? | ? |
| Total | 100 | 100 |

During the study we retrospectively determine the number of smokers and non-smokers in both the case and control group.

Our table might then look like this:

Figure 4 Contingency table at end of retrospective study of smoking and lung cancer

| | Cases | Controls |
|---|---|---|
| Smokers | 90 | 40 |
| Non-smokers | 10 | 60 |
| Total | 100 | 100 |

Now we can compare exposure rates for cases and controls:

Case exposure rate $= \frac{a}{a+c} = \frac{90}{100} = 90\%$

Control exposure rate $= \frac{b}{b+d} = \frac{40}{100} = 40\%$

If necessary we can determine the statistical significance of the result using the chi-square test (Unit 7).

**Cohort studies**

Study design

Prospective or cohort studies begin with the exposed group (a+b) and non-exposed group (c+d) to give the following position at the start of the study.

Figure 5 Contingency table at beginning of prospective study

| | | Disease | | |
|---|---|---|---|---|
| | | Present | Absent | Total |
| Exposure | Present | ? | ? | a+b |
| | Absent | ? | ? | c+d |

The participants are then followed forward in time so that eventually they fall into the disease or non-disease columns, enabling the 2 x 2 table in Figure 1 to be completed prospectively.

Analysis

The analysis is performed by comparing the rate of disease occurrence (incidence) between exposed and non-exposed groups.

Incidence in exposed group $= \frac{a}{a+b}$

Incidence in non-exposed group $= \frac{c}{c+d}$

Let us look at an example.

Example

Consider a cohort of 2000 persons of whom 800 are smokers and 1200 are non-smokers. At the beginning of the study our 2 x 2 table is as follows.

Figure 6 Contingency table at beginning of prospective study of smoking and lung cancer

| | Lung cancer | No lung cancer | Totals |
|---|---|---|---|
| Smokers | ? | ? | 800 |
| Non-smokers | ? | ? | 1200 |

The entire cohort is followed for twenty years and 100 develop lung cancer, 90 of whom are smokers and 10 are not. We now get a 2 x 2 table like this:

Figure 7 Contingency table at end of retrospective study of smoking and lung cancer

| | Lung cancer | No lung cancer | Total |
|---|---|---|---|
| Smokers | 90 | 710 | 800 |
| Non-smokers | 10 | 1190 | 1200 |
| Total | 100 | 1900 | 2000 |

We next compare incidence rates for smokers and non-smokers:

Smokers: $\frac{a}{a+b} = \frac{90}{800} =$ 112.5 per 1000

Non-smokers: $\frac{c}{c+d} = \frac{10}{1200} =$ 8.3 per 1000

If necessary we can determine the statistical significance of the result using the chi-square test (see Unit 7).

**RELATIVE RISK**

Relative risk is a measure of the strength or magnitude of the association between a factor and a certain outcome. Therefore a high relative risk suggests aetiology or causation.

Relative risk is defined as the incidence rate among exposed divided by the incidence rate among non-exposed (see Unit 2, section 2.2 if you need to revise measures of risk).

Prospective studies

In a prospective study we can determine incidence rates directly in those exposed and those not exposed. Therefore we can calculate relative risk as the ratio of the two incidences.

$$\text{Relative risk} = \frac{a/(a+b)}{c/(c+d)}$$

In the example in Figure 7 the relative risk is

$$\frac{90/800}{10/1200} = 13.5$$

Retrospective studies

In retrospective studies it is not possible to determine incidence rates for the exposed or non-exposed groups. Instead of relative risk we calculate a statistic called the odds ratio.

Odds ratio

When the following conditions are met the odds ratio gives a close approximation to the relative risk:

- The disease has a low incidence in the general population (this is true for many chronic diseases).
- The control group is representative of the general population with respect to frequency of the attribute.

Using the symbols in Figure 1:

$$\text{Odds ratio} = \frac{ad}{bc}$$

In the example in Figure 4 the odds ratio is:

$$\frac{90 \times 60}{40 \times 10} = 13.5$$

and provides an estimate of the relative risk for smokers.

*Exercise 1*

*Using the data on pregnancy and anaemia in Figure 8, calculate the odds ratio and relative risk. What does the odds ratio tell you about the relationship between pregnancy and anaemia?*

Figure 8 Contingency table showing relationship between pregnancy and anaemia among women

| Anaemia | No anaemia | Total | |
|---|---|---|---|
| Pregnant | 600 (30%) | 1400 (70%) | 2,000 (100%) |
| Not pregnant | 160 (2%) | 7840 (98%) | 8,000 (100%) |
| Total | 760 | 9240 | 10,000 |

Figure 9 Episodes of ARI and pneumonia in 165 children followed up for a year (nutritional state determined at beginning of study).

| | Malnourished | | | Well-nourished | | |
|---|---|---|---|---|---|---|
| Age in months | No. children | No. ARI attacks | No. pneumonia attacks | No. children | No. ARI attacks | No. pneumonia attacks |
| 0-12 | 30 | 127 | 21 | 22 | 103 | 3 |
| 13-36 | 42 | 212 | 14 | 39 | 210 | 1 |
| 37-40 | 17 | 95 | 9 | 15 | 67 | 2 |
| Totals | 89 | 434 | 44 | 76 | 380 | 6 |

*Exercise 2* *Calculate and tabulate the age-specific incidence (attack rates) in malnourished and well-nourished children. What do these rates indicate?*

*Exercise 3* *Calculate the prevalence of malnutrition.*

*Exercise 4* *Calculate the relative risk of ARI and pneumonia among malnourished children. What do these results mean?*

Answer 1

$$\text{Odds ratio} = \frac{ad}{bc} = \frac{4704,000}{224,000} = 21$$

The odds ratio indicates that the odds in favour of having anaemia are 21 times higher among pregnant than among non-pregnant women, or that the odds in favour of being pregnant are 21 times higher in anaemic than among non-anaemic women.

$$\text{Relative risk} = \frac{600/2000}{100/8000} = 15$$

or, since the percentages are already calculated in the table,

$$\frac{30\%}{2\%} = 15$$

Answer 2

Age-specific incidence (attack) rates

| Age in months | Malnourished ARI | Malnourished Pneumonia | Well-nourished ARI | Well-nourished Pneumonia |
|---|---|---|---|---|
| 0-12 | 4.23 | 0.7 | 4.68 | 0.14 |
| 13-36 | 5.05 | 0.3 | 5.38 | 0.03 |
| 37-60 | 5.59 | 0.53 | 4.46 | 0.13 |
| Overall (crude) rate | 4.88 | 0.49 | 5.0 | 0.08 |

There is no difference in the ARI attack rates between well-nourished and malnourished children, but there is a massive difference for pneumonia attack rates. There is no definite age trend.

Answer 3

$$\text{The prevalence of malnutrition} = \frac{89 \times 100\%}{89+76} = 53.9\%$$

Answer 4

The relative risk is calculated by the attack rate among malnourished divided by the attack rate among normal children.

$$\text{For ARI the relative risk} = \frac{4.88}{5.0} = 0.98$$

$$\text{For pneumonia the relative risk} = \frac{0.49}{0.08} = 6.13$$

These figures mean that malnourished children are 0.98 times less likely than normal children to get ARI, but 6.13 times more likely to get pneumonia.

---

## 6.3 INTERPRETING ASSOCIATIONS

As we mentioned earlier, the outcome of an analytical study is the conclusion that a disease and its suspected cause are, or are not, associated. If the analysis of a contingency table reveals an association, you will need to ask yourself three questions.

Questions about associations

1 Is the observed association the result of bias in the investigation?
2 Is the observed association the result of chance?
3 If the association is a real one, is it a cause and effect relationship?

Let us see how we might answer these questions.

### BIAS

The first step must be an enquiry into the possibility of bias in the study. Bias is a systematic error resulting in over- or underestimation of the strength of the association. Associations which are produced by flaws in the design or execution of a study that result in bias, are known as artifactual or spurious associations. They are not true associations. Steps must be taken at the planning phase to reduce the possibility of bias. Refer back to Unit 3, section 3.3 (advantages and disadvantages of case-control and cohort studies) and section 3.4 (selection of controls).

Artificial associations

Diagnosing cases

The validity of a case-control or cohort study will depend on the accuracy with which the subjects are assigned to each of the four categories in the contingency able (a, b, c and d in Figure 1). Misclassification may occur because of over- or under-diagnosis. Are you sure that the disease was in fact absent in all the controls? It may be easy to classify people with and without a well-defined disease such as lung cancer. But how certain can you be that you have correctly classified subjects with a disease like tuberculosis. This will depend on the accuracy of your diagnostic technique (see section 3.5).

Diagnosing exposure

Similarly, exposure to a suspected risk factor may be difficult to determine, or it may be intermittently present, for example oral contraceptive use.

### CHANCE

We next need to find out whether the association is a chance finding due to the fact that we have investigated a sample rather than the entire population (see Unit 3). To do this we need to perform a significance test. The chi-square test is the usual significance test for analysing contingency tables (see Unit 7 for information on significance tests).

Significance tests

Irrespective of the outcome of significance tests, it is often necessary to confirm that the observations can be repeated using other sets of data.

## CAUSAL AND NON-CAUSAL RELATIONSHIPS

If we eliminate the possibility that the association is due to chance or to a faulty study design causing bias, it is likely that the association is a real one. However it may not be a causal association. There are three possibilities.

- The association is causal: the exposure causes the disease (see section 6.4).
- The association is non-causal:
  - the disease may cause the exposure;
  - both disease and exposure are associated with a third, unknown factor, called a confounding variable.

Non-causal associations

Let us look at an example to illustrate this third possibility.

Example

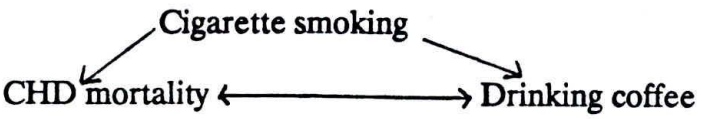
A statistically significant association was found between mortality rates for coronary heart disease (CHD) and coffee consumption. People who drink a lot of coffee also tend to be cigarette smokers and cigarette smoking is strongly associated with CHD mortality. Using a cross-classification technique (see Unit 5, section 5.7), when cigarette consumption is held constant the effect of coffee drinking disappears.

Figure 10 Cross-tabulation of CHD mortality rates according to cigarette smoking and coffee consumption

CHD Mortality rates in males
55-64 years (deaths/1000 per year)

| Coffee (cups/day) | Cigarettes smoked/day 0 | 20-40 | 60+ | All |
|---|---|---|---|---|
| 0 | 4 | 9 | 15 | 6 |
| 1-5 | 6 | 10 | 13 | 8 |
| 6+ | 5 | 9 | 16 | 12 |
| All | 4 | 10 | 15 | |

Thus the association between coffee drinking and CHD is non-causal, mediated by the confounding variable, cigarette smoking. If coffee drinking is varied independently of cigarette consumption, CHD mortality rates are unchanged.



## CONFOUNDING VARIABLES

Confounding variables are usual in associations between diseases and their causes. For example, there are very few causes or diseases which are not influenced by age or sex. There are two ways of minimising this problem (see also Unit 3, section 3.4).

Matching

- An investigator may choose individuals for both the study and control groups who are similar (matched) with respect to a confounding variable, for example age. Therefore, for every person aged 30 years in the study group, the investigator will choose one 30 year old for the control group.

Stratification

- The variable suspected of being a confounding factor can be held constant by cross-tabulation or stratification. This was the technique used in our example of coffee consumption, cigarette smoking and lung cancer (see Figure 10). The investigator separates into groups or strata those who possess different levels of the confounding variable. Groups with the same level of confounding variable are then compared to see if the differences persist.

  For example, if age is a potential confounding variable, you would subdivide the groups by age into several categories. Then you would compare the study and control groups in each age classification to determine whether differences persist when the effects of age are eliminated.

---

## 6.4 CAUSAL RELATIONSHIPS

In section 6.3 we discussed the various kinds of non-causal association. In this final section we shall consider the various types of causal relationships, and how we can establish causation.

### EXAMPLES OF CAUSAL RELATIONSHIPS

Let us look at some examples of causal relationships to see how complex they can be.

One cause

1 A ——> B

A straightforward relationship in which A causes B (i.e. the cause always produces the disease) seldom occurs.

Two causes

2 A + C ——> B

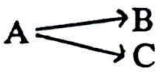
Both A and C are required to produce the disease. This is probably the correct model for many diseases. For example, there is little doubt that smoking (A) causes lung cancer (B), but not all smokers get lung cancer; some other factor (C) may mediate or influence the development of the disease.

Two independent causes

3 


Either A or C acting alone is capable of causing B. For example hepatitis (B) may result from viral infection (A) or chemical agents (C). This model is important because it establishes that a disease may be caused in different ways.

Two outcomes

4 

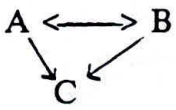

A may cause either B or C (one cause can produce different outcomes). For example, asbestos dust exposure (A) can cause chronic lung disease (B) or lung cancer (C).

Chain of events

5 A —> B —> C

Disease C is caused by a chain of events: A and B are incidental steps leading to the disease. For example tooth decay or gum disease (A) may lead to inadequate nutrition (B) which may lead to vitamin deficiency disease (C).

Interrelated causes

6 


Each factor may be either the cause or the result (outcome) of other factors. This chain of events may begin at any point of the triangle. For example protein or other dietary deficiency (A) may lead to, or be the result of, intestinal malabsorption of nutrients (B), which in turn may lead to, or be the cause of, malnutrition or generalised debilitation (C).

Web of causation

As more factors or variables are implicated in the disease process, the notion of causation becomes confusing. The number of variables soon spreads to include a large diverse body of information which is called 'the web of causation'. This implies that a disease has many causes and results from the relative importance of, and relationship between, the different causes.

## ESTABLISHING CAUSATION

Five criteria have generally been adopted as a test of causation.

Consistency

- The consistency of the association. Different studies should result in the same association even though they may use different designs or be conducted on different populations.

Strength

- The strength of the association. The greater the relative risk the more likely it is that the association is causal. The likelihood that the exposure is causal increases if a dose-response gradient can be demonstrated. Dose-response gradients can be recorded as the degree (number of cigarettes smoked daily) or duration of exposure to the supposed causative factor.

Specificity

- The specificity of the association. This is the degree to which one particular exposure produces one specific disease. Although smoking plays a role in many diseases the effect is greatest with lung cancer. Thus it is more likely to play a more direct and causal role in lung cancer than in other diseases.

Time

- The temporal relationship of the association. Exposure to the factor must precede development of the disease.

Coherence

- The coherence of the association. A causal association should be consistent with and supported by available biological data, such as laboratory experiments on animal models.

# UNIT 7 DATA ANALYSIS: STATISTICAL TESTS

## UNIT 7: OBJECTIVES

Study with this unit will enable you to:

- Interpret statements of statistical significance and explain when and why significance tests are used.
- Calculate standard errors of means and proportions, and use them to compute 95% confidence limits and to test for statistical significance.
- Explain the use of a chi-square test and perform the test with the aid of reference material.
- Explain the information provided by a correlation coefficient and a regression equation, and calculate these statistics with the aid of reference material.

---

# UNIT 7: CONTENTS

## 7.1 INTRODUCTION

Why use statistics?

We have discussed in Units 5 and 6 how we can analyse data to show frequency distributions and associations between variables. For many studies this may be sufficient. These analytical methods may give you all the information you need in order to achieve the objectives of your study. But sometimes you can get more useful information out of your data by doing simple statistical tests. Let us look at three examples to illustrate how you can use statistics and why you need to use statistics.

Estimating population values

Imagine that you have recorded the diastolic blood pressure in a random sample of 100 men from several villages in your district. You summarise the results by calculating the mean and standard deviation (see Unit 5). You would like to generalise from the results of your sample back to the entire population. That is, you would like to estimate the mean blood pressure in the population from which the sample was taken. You can only do this using statistical techniques. In this case we calculate 95% confidence limits (see section 7.2).

Measuring the difference between groups

For our second example, you have obtained data on the live birthweights recorded in your hospital for 1986 and for 1976. You have drawn up frequency distributions, and calculated means and standard deviations for the two sets of data (Unit 5). You think that there has been an increase in mean birthweight at the hospital over this ten-year period. How can you be sure that this difference is a real one and not just due to random variation? A significance test (section 7.3) will show whether the observed difference between the mean values of the two sets of data is too large to be explained by random variation (chance) alone.

Studying assocations

Here is a third example. Suppose that you have analysed the results of your study and found an association between pregnancy and anaemia. If you had studied the entire population, you could be sure that this association was a real one. But you have examined only a small sample of pregnant women. Even though you have chosen your sample well (see Unit 3, section, 3.4), there is still the possibility that the result could be due to chance. In order to exclude this possibility you would need to do a significance test. This will tell you how likely it is that the association is a real or significant one, rather than a chance finding. (Note that it will not tell you if the association is a cause and effect one - see Unit 6).

Uses of statistics

We have therefore identified three uses of statistics:

- Estimating population values.
- Measuring the difference between groups.
- Studying associations.

We need statistics for two reasons:

Variability

- Variation occurs in observational and experimental data.

Sampling

- Most studies are done on samples of populations because it is not practical to study the entire population.

If you studied the whole population, for example in a census, then any differences you found would be true differences, and the rates you observed would be true rates. This is because your conclusions are not based on partial samples of the population and there would be no sampling error (see Unit 3, section 3.4). Therefore, you would need only to measure the association or differences which you found and report them.

However, when dealing with samples, we have to try to draw conclusions about the population on the basis of the information obtained from the sample using what are known as sample statistics.

Sample statistics

The use of sample statistics rests on the assumption that the sample is representative of the population, that is, the sample has been obtained by random sampling. This unit is concerned with two uses of sample statistics:

- Estimating population values.
- Answering questions about population values in the form of testing hypotheses.

---

## 7.2 ESTIMATING POPULATION VALUES

### STANDARD ERROR AND CONFIDENCE LIMITS

Probability sampling

In Unit 3, section 3.4 we discussed the importance of probability sampling - that when you draw a sample from a population you ensure that every member of the population has a known chance of being included in the sample. It is then possible to estimate the probable findings in the parent population from which the sample was drawn.

Variation between samples

Samples that are drawn from the parent population will show chance variation from one to another, and this variation may be small or large. What determines the variation between samples? There are two factors.

Variation

- The variation depends partly on the amount of variation in the population from which the samples were drawn. For example a series of samples of the body temperature of healthy people would show very little variation from one another, while samples of systolic blood pressure would show considerable variation.

Sample size

- The variation between samples depends on the size of the sample. A small sample is a much less certain guide to the population from which it was drawn than a large sample.

Therefore variation depends on the variation in the population and the size of the sample. As we do not know the variation in the population we use as an estimate of it the variation in the sample as expressed by the standard deviation.

Standard error of the error

If we divide the standard deviation by the square root of the number of observations in a sample we have the standard error of the mean (SEM).

$$SEM = \frac{SD}{\sqrt{n}}$$

The standard error of the mean therefore describes the reliability of a sample mean in indicating the true mean of the whole population.

Example

Now for an example. A study of diastolic blood pressure was made on a random sample of 72 factory workers and 48 farmers.

Figure 1 Mean diastolic blood pressure in mmHg of factory workers and farmers

| | No. | Mean diastolic blood pressure | Standard deviation |
|---|---|---|---|
| Factory workers | 72 | 88 | 4.5 |
| Farmers | 48 | 79 | 4.2 |

The standard errors of the two mean blood pressures are calculated as follows:

Factory workers: $SEM = \frac{4.5}{\sqrt{72}} = 0.53$

Farmers: $SEM = \frac{4.2}{\sqrt{48}} = 0.61$

Now, how can we use the standard error of the mean to generalise from a sample back to the population from which it came? This is done using a useful device called confidence limits.

Confidence limits

Confidence limits are so called because they are determined in accordance with a specified or conventional level of confidence or probability that these limits will in fact include the population value or parameter (e.g. mean) being estimated. The interval between the

Confidence interval

confidence limits is known as the confidence interval or confidence range. Thus 95% confidence limits estimate with 95% reliability the range within which the larger population's average lies (see also Unit 3, section 3.4). The level of confidence is conventionally set at 95% or 0.95. To calculate 95% confidence limits we need three pieces of data: the sample standard deviation, the sample size and the sample mean.

Calculation

The first step is to calculate the standard error of the mean. From this, 95% confidence limits can be calculated by the formula: sample mean $\pm$ 2 x SEM.

Example

To go back to our previous example, SEM for factory workers was 0.53.

$\therefore$ 95% confidence limits $= 88 + (2 \times 0.53) = 89.06$

$= 88 - (2 \times 0.53) = 86.94$

Therefore we can say that there is only a 5% chance or less that the mean of the total population lies outside the range of 86.96 to 89.04 mmHg.

*Exercise 1*

*A count of malaria parasites in 100 fields gave a mean of 33 parasites per field, standard deviation 11.1. What are the 95% confidence limits for the mean of the population from which this sample count of parasites was drawn?*

The answer to this exercise is on the next page.

The concepts of standard errors and confidence limits can also be used to determine whether a statistically significant difference exists between two samples (see section 7.5).

**STANDARD ERROR OF PERCENTAGES**

In the same way that we can calculate the standard error of a mean, so we can calculate the standard error of a percentage. Again, both the size of the sample and the amount of variation of the population from which it is drawn will affect the size of the standard error.

Example

Here is an example. Examination of hospital records of burns cases over a ten-year period showed that of 120 cases aged 60 years or over, 73 (60.8%) were in women and 47 (39.2%) were in men. If p represents one of the percentages and 100 - p represents the other then the standard error of the percentage is obtained by multiplying them together, dividing the result by the number in the sample and then taking the square root.

$$\text{SE percentage} = \sqrt{\frac{p\,(100 - p)}{n}}$$

In our example,

$$\text{SE percentage} = \sqrt{\frac{60.8 \times 39.2}{120}} = 4.46$$

We obtain 95% confidence limits in the same way as for the standard error of the mean:

60.8 + (2 x 4.46) = 51.88 and 69.72 (women)
39.2 + (2 x 4.46) = 30.28 and 48.12 (men)

Therefore there is a probability of 95% that the percentage of men and women in the population from which the sample came fall within these confidence limits.

Answer 1

The first step is to calculate the standard error of the mean.

$$\text{SEM} = \frac{SD}{\sqrt{n}} = \frac{11.1}{\sqrt{100}} = 1.11$$

95% confidence limits = 33 + (2 x 1.11) = 35.22
33 - (2 x 1.11) = 30.78

Therefore, there is a 95% certainty that the mean number of counts in the large population from which the sample was drawn lies between 30.78 and 35.22.

---

## 7.3 TESTING HYPOTHESES

### WHAT IS STATISTICAL SIGNIFICANCE ?

Because most epidemiological investigations are conducted on only a sample of the population, we are confronted with the question of whether we would obtain similar results if we had studied the entire population or whether chance selection (sampling variation) has produced unusual results. We can answer this question using a test of significance.

Meaning

The term statistically significant is often found in the medical literature but its meaning is still widely misunderstood. Let us consider a clinical example to illustrate how statistical significance can help us to interpret comparison results.

Example

In a small series of patients, a clinician finds that the mean response to treatment is greater for drug A than drug B. Now he would obviously like to know if the difference he has observed in his small series of patients would also be true for all such patients. In other words, he wants to know if the observed difference is more than merely sampling error. This assessment can be made with a statistical test.

Reasons for difference

The possible reasons which could account for the differences observed in response to drugs A and B are:

1 The difference may be due to random variation in response (sampling variation).
2 The difference may be due to bias: some factor such as age of patients has not been controlled in any way.
3 The difference may be a true one: drug A actually could be superior to drug B.

Bias

We can only conclude that the difference is a real one after we have ruled out 1 and 2 as possibilities. To rule out reason 2 we must make sure that the study design does not lead to biased results (see Unit 3).

Random variation

To rule out reason 1 we test for statistical significance. If the test shows that the observed difference is too large to be explained by random variation (chance) alone, we state that the difference is statistically significant. We therefore conclude that we would have found this difference between the two drugs if we had studied the whole population of these patients.

### COMMON USES OF SIGNIFICANCE TESTS

Statistical techniques as applied to medical studies are commonly used in two situations:

- To measure the difference between groups - are the differences statistically significant? That is, are the observed differences between groups likely to reflect a true difference in the larger population?

between groups likely to reflect a true difference in the larger population?

Associations

- To study associations: do two characteristics occur together more often than would be expected by chance?

## WHEN SIGNIFICANCE TESTS ARE NOT REQUIRED

Do not do a significance test unless you are sure it is necessary and will help you to interpret your results. Significance tests are not needed:

No sampling

- When entire populations are studied (for example census data or death certificate data) because there is then no sampling error.

No probability sampling

- When probability sampling (see Unit 3, section 3.4) is not used, for example in some simple descriptive surveys and programme reviews.

No practical difference

- When the results of significance tests will make no difference in practice. For example for practical purposes it may be enough to know that most cases of obstructed labour occur in several very remote villages without worrying about testing whether the associations have occurred by chance.

## HOW DOES A SIGNIFICANCE TEST WORK?

Null hypothesis

Underlying all statistical tests is a 'null hypothesis' see also Unit 3, section 3.2). The null hypothesis states that there is no difference in population parameters among the groups being compared. In other words, the null hypotheses is consistent with the idea that the observed difference is simply the result of random variation in the data. To decide whether we should accept or reject the null hypothesis we calculate a test statistic and compare it with a 'critical value' obtained from a set of statistical tables. When the test statistic exceeds the critical value we reject the null hypothesis and say that the difference is statistically significant.

Testing the null hypothesis

Steps in testing for significance

The procedure for significance testing is summarised below. There are many types of significance test, each appropriate for different kinds of data, but these basic principles apply to the procedure as a whole, and are not the details of how to perform specific tests. We will illustrate each step by referring to the experience of Dr A who is investigating episodes of diarrhoea.

Hypothesis

1 **State hypothesis**

Develop the study question: an association exists between factors or a difference exists between groups in the general population.

*Dr A was concerned about the incidence of diarrhoea in a nearby village. He thought that the infections might be related to the village water supply so he decided to investigate his theory by doing a simple study.*

Null hypothesis

2 **Formulate null hypothesis**

Reverse the hypothesis: no association between factors or difference between groups exists in the general population.

*Dr A formulated his null hypothesis: 'There was no association between the source of water and diarrhoeal episodes'.*

Significance level

3 **Decide significance level**

⩽5% unless otherwise indicated (see next page).

*Dr A decided to choose the conventional significance level of 5%.*

Data

4 **Collect data**

Determine whether an association between factors or a difference between groups exists in the data collected from samples of the larger population.

*Dr A went to the village and recorded the water supply of 124 households. He reviewed the village health centre's morbidity records for the previous three months and identified household members with a history of diarrhoeal episodes. He summarised the data he collected in a contingency table (Figure 2).*

Figure 2 Three-month history of diarrhoea episodes

| | No. of households according to water supply | | | Total |
|---|---|---|---|---|
| | River | Well | Tap | |
| No diarrhoea | 39 | 14 | 12 | 65 |
| Diarrhoea episodes | 49 | 6 | 4 | 59 |
| Total | 88 | 20 | 16 | 124 |

Significance test

5 **Apply statistical significance test**

Determine the probability of obtaining the observed data if the null hypothesis were true, i.e., choose and apply the correct statistical significance test.

*He thought that his data indicated that there was an association between water supply and diarrhoeal episodes. But he wanted to be sure that the result was not just due to chance. Therefore he applied a statistical test. (He actually used a chi-square test: see section 7.4).*

Verdict

6 **Reject or fail to reject the null hypothesis**

Reject the null hypothesis and accept by elimination the study hypothesis if the significance level is reached. Fail to reject the null hypothesis if the observed data has more than a 5% probability of occurring by chance.

*The result of the test was statistically significant at the 5% level. That is, the association had less than a 5% probability of occurring by chance alone. Dr A therefore rejected the null hypothesis that there was no association. He concluded that there was indeed an association between diarrhoeal episodes and source of water in the village. (Note: to establish a causal relationship he would have to make further investigations. See Unit 6).*

**Significance levels**

Definition

Any decision to reject the null hypothesis carries with it a certain risk of being wrong. This risk is called the significance level of the test. If we test at the 5% significance level, we are taking a 5% chance of rejecting the null hypothesis when it is true. Naturally we want the significance level to be small, and in most studies the 5% level is used.

*P* values

The lowest significance level at which the null hypothesis could be rejected is called the *P* value. This expresses the probability that a difference as large as we have observed would occur by chance alone. A *P* value of 0.05 indicates a 5% probability of achieving the observed difference or association if the null hypothesis is true. The level of significance which is selected must be clearly stated otherwise the statement that the results are 'statistically significant' is worthless.

It is usual to indicate your conclusion and significance level simultaneously in the format:
'statistically significant ($P<0.05$)' or
'not statistically significant ($P>0.05$)'.

**INTERPRETATION OF SIGNIFICANCE**

It is important to remember that a test of significance always refers to a null hypothesis. It answers the question: 'Is chance or sampling variation a likely explanation of the discrepancy between a sample result and the corresponding null hypothesis population value?' We can summarise the relationships between the meaning of 'statistically significant' and 'not statistically significant', and the null hypothesis as follows.

| | | | |
|---|---|---|---|
| Statistically significant | = Reject null hypothesis | = Sample value not compatible with null hypothesis value | = Sampling variation an unlikely explanation of discrepancy between null hypothesis and sample values |
| Not statistically significant | = Do not reject null hypothesis | = Sample value compatible with null hypothesis value | = Sampling variation is a likely explanation of discrepancy between null hypothesis & sample values |

Non-significance

A statistically significant difference is one that cannot be accounted for by chance alone. The converse is not true - that is a difference that is not statistically significant is not necessarily attributable to chance alone. In the case of a non-significant difference the sample size is very important, because small samples are associated with large sampling errors which may lead to a non-significant test even when the observed difference is caused by a real effect. Therefore a

result that is non-significant should be regarded as inconclusive or 'not proved' rather than an indication of no effect.

Clinical importance

One further word of warning: do not equate 'statistical significance' with 'medically important'. For example, if large samples are studied, very small differences that have no clinical importance may turn out to be statistically significant.

Do not judge the practical implications of any finding on statistical grounds alone.

*Exercise 2*

*The mean birthweight of first-born infants of 45 women who smoked twenty cigarettes per day during pregnancy was 200 grams lower than that of the first-born infants of 39 women who never smoked. The difference was statistically significant at the 5% level ($P<0.05$). What do you conclude?*

.....

.....

.....

*Exercise 3*

*In a study of 100 cases of a disease and 100 controls, you find that the difference with respect to a possible aetiological agent is not statistically significant. What do you conclude?*

.....

.....

.....

Answer 2

You should conclude that the difference observed between mean birthweights was too large to have occurred by chance alone. (Note: statistical significance indicates that smoking during pregnancy is associated with low birthweight. It does not establish causation).

Answer 3

You should conclude that the difference may be the result of sampling variation. (Note: statistical non-significance does not indicate that there is no association of the factor with the disease).

**CHOOSING A SIGNIFICANCE TEST**

The selection of a specific statistical test depends on the following factors.

- Type of data (nominal, ordinal or continuous data - see Unit 5).
- Number of groups to be compared (two or more than two).
- Have study and control groups been matched (paired)?
- Is a one-tailed or two-tailed significance test needed?

One-tailed or two-tailed tests

With a statistically significant one-sided test result (upper tail) we can infer that the true population value is above that specified by the null hypothesis. With a two-sided test we can infer that the true

population value is either above or below the null hypothesis value. When in doubt, use a two-tailed test.

Some common statistical tests

In the remaining sections of this unit we will consider the three types of statistical methods which you are likely to find most useful.

- The chi-square test (section 7.4). This is one of the most widely used tests in epidemiological studies for investigating associations.
- Comparison of two means or two percentages using the standard error (section 7.5). This is a simple technique for comparing two groups using the concept of confidence limits.
- Correlation and regression (section 7.6). These are two methods used to describe the relationship between two variables depicted in a scatter diagram.

If these tests are not appropriate for the data you wish to analyse you should consult your supervisor, if you have one, or a statistics textbook, or an experienced epidemiologist or statistician.

---

## 7.4 THE CHI-SQUARE TEST

The chi-square ($\chi^2$) test determines whether the **observed** frequencies of individuals with given characteristics differ significantly from the frequencies which would be **expected** under some theory or hypothesis.

Uses

It is used for the analysis of contingency tables (see Unit 5) which are tables of frequencies showing how the total frequency is distributed among the cells of the table. Contingency tables are usually constructed for the purpose of showing the relationship between two variables. The $\chi^2$ test is a means of testing the null hypothesis that the two variables are independent, i.e. no relationship exists between them.

Conditions for using $\chi^2$

If you are thinking of using a $\chi^2$ test, remember:

- The $\chi^2$ test is only used on actual numbers or counts of occurrences in a category, and not on percentages, proportions or measured quantities, means of observations or other derived statistics.
- The misapplication of the $\chi^2$ test to data other than counts represents one of the most common errors committed by investigators inexperienced in statistical methods.
- The other major condition for its use is that in 2 x 2 contingency tables $\chi^2$ cannot be used if the total in the table is less than 20, or if the smallest expected (not observed) value is less than 5.

Study advice

Reference 3 describes how to do a $\chi^2$ test. Read it carefully and then try the following exercises. Set out all the stages in your calculation so that if you make a mistake you can easily find the source of your error by checking with our answers. There are also two exercises in Reference 3 (pages 463 and 514) which you could try for additional practice!

*Exercise 4*

*Using the data in Figure 3, calculate $\chi^2$. Is the value statistically significant at the 0.05 level? What do you conclude?*

Figure 3 Children aged 2-9 years with and without splenic enlargement by sex

| | With enlarged spleen | Without enlarged spleen | Total |
|---|---|---|---|
| Males | 21 | 59 | 80 |
| Females | 37 | 63 | 100 |
| Total | 58 | 122 | 180 |

*Exercise 5* *Let us return to the data collected by Dr A which are repeated in Figure 4 below. Analyse the results using $\chi^2$. What do you conclude from your analysis?*

Figure 4 Three-month history of diarrhoea episodes

| | No. of households according to water supply | | | Total |
|---|---|---|---|---|
| | River | Well | Tap | |
| No diarrhoea | 39 | 14 | 12 | 65 |
| Diarrhoea episodes | 49 | 6 | 4 | 59 |
| Total | 88 | 20 | 16 | 124 |

Answer 4

As this is a simple 2 x 2 table we can use the quick formula for $\chi^2$ described on page 513 of Reference 1.

$$\chi^2 = \frac{(21 \times 63 - 59 \times 37)^2 \times 180}{80 + 100 + 58 + 122} = 2.352$$

Looking up this result in the $\chi^2$ table (shown in Figure 14.3 on page 463 of Reference 3) we find that at the 0.05 significance level $\chi^2$ with one degree of freedom = 3.841. Our $\chi^2$ value is less than this, therefore we conclude that the difference is not significant. This means that the different rates of splenic enlargement in boys and girls could have occurred by chance.

Answer 5

The stages in your calculation should follow those in Reference 3 page 462. Figure 4 shows the observed values. Expected values are calculated as follows:

| | River | Well | Tap |
|---|---|---|---|
| No diarrhoea (A) | $\frac{88 \times 65}{124}$ | $\frac{20 \times 65}{124}$ | $\frac{16 \times 65}{124}$ |
| Diarrhoea (B) | $\frac{88 \times 59}{124}$ | $\frac{20 \times 59}{124}$ | $\frac{16 \times 59}{124}$ |

$\chi^2$ can then be calculated by the formula

$$\chi^2 = \frac{(\text{observed} - \text{expected})}{\text{expected}} \text{ or } \frac{(O-E)^2}{E}$$

| | | River | Well | Tap | Total |
|---|---|---|---|---|---|
| Expected No. | A | 46.13 | 10.48 | 8.39 | 65 |
| | B | 41.87 | 9.52 | 7.61 | 69 |
| O-E | A | -7.13 | 3.52 | 3.61 | |
| | B | 7.13 | -3.52 | -3.61 | |
| $\frac{(O-E)^2}{E}$ | A | 1.102 | 1.182 | 1.553 | 3.837 |
| | B | 1.214 | 1.302 | 1.712 | 4.228 |

$\chi^2 = 3.837 + 4.228 = 8.065$

The number of degrees of freedom = (number of columns -1) x (number of rows -1) = 1 x 2 = 2
From the table of $\chi^2$ distribution, $\chi^2$ at 0.05 with 2 degrees of freedom = 5.991. As our value for $\chi^2$ is greater than this, the result is statistically significant at $P < 0.05$. (i.e. we reject the null hypothesis at the 5% level). Therefore these data indicate an association between diarrhoeal episodes and source of water in the village. Note that further investigations would be needed in order to establish a causal relationship.

## 7.5 SIGNIFICANCE AND STANDARD ERROR

Comparing means

We saw earlier in this section that the mean of a sample has a standard error, and a mean that departs by more than twice its standard error from the population mean would be expected by chance in only about 5% of samples. Similarly, the difference between the means of two samples has a standard error, and we can use the standard error of the difference between means to test the null hypothesis that there is no difference between the means of two samples.

Large samples only

Note that we can only use this method to compare large samples, preferably when there are more than 60 observations. It should not be used to compare samples of 30 observations or less. In this case you need to use a t-test. Please consult your supervisor or a statistics textbook if you need to do a t-test.

t-tests

### STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO MEANS

To explain how this test is done, we will go back to the example of the diastolic blood pressure in factory workers and farmers from page 201. The data are repeated below.

Figure 5 Mean diastolic blood pressure in mmHg of factory workers and farmers

| | No. | Mean diastolic blood pressure | Standard deviation |
|---|---|---|---|
| Factory workers | 72 | 88 | 4.5 |
| Farmers | 48 | 79 | 4.2 |

Calculation

To compare the means of the two groups of workers we perform the following steps:

- Erect the null hypothesis that there is no significant difference between the mean blood pressures of the two groups.
- Calculate the standard error of the difference. To do this for two samples, 1 and 2, we need to know only the standard deviation of each sample ($SD_1$ and $SD_2$) and the number in each sample ($n_1$ and $n_2$)

Then, SE difference = $\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$

Using the data from Figure 5,

SE difference = $\sqrt{\frac{4.5^2}{72} + \frac{4.2^2}{48}} = 0.81$ mmHg

- Calculate the difference between the means. In this case 88 - 79 = 9 mmHg
- Divide the difference between means by the standard error of the difference, i.e. $\frac{9}{0.81} = 11.1$

This indicates that the difference between the means is 11.1 times greater than the standard error of the difference.

Determining significance

In order to determine whether the difference is statistically significant we refer to a probability table like the one shown in Figure 6.

Figure 6 Probability related to multiples of standard deviations (or standard errors) for a normal distribution.

| Number of standard deviations | Probability of observations showing at least as large a deviation from the population mean |
|---|---|
| 0.674 | 0.5 |
| 1.645 | 0.1 |
| 1.960 | 0.05 |
| 2.326 | 0.02 |
| 2.576 | 0.01 |
| 3.291 | 0.001 |

By reference to the table we can see that our value of 11.1 exceeds that of 3.291 which corresponds to a probability of 0.001 ($P < 0.001$). The possibility of a difference of 11.1 standard errors occurring by chance is thus exceedingly low, and so the null hypothesis that these two samples come from the same population of observations is extremely unlikely.

*Exercise 6*

*In one group of 62 patients with iron deficiency anaemia the haemoglobin level was 12.2 g/dl, standard deviation 1.8 g/dl. In another group of 35 patients it was 10.9 g/dl, standard deviation 2.1 g/dl. What is the standard error of the difference between the two means? What is the significance of the difference? The answers are given on page 218.*

**STANDARD ERROR OF THE DIFFERENCE BETWEEN PERCENTAGES**

We can apply a similar technique to comparing two percentages. On page 202 we described how to calculate the standard error of a percentage. We will now continue with the example we used then to illustrate how two groups can be compared.

Example

In our original example, examination of the hospital records of burns cases over a ten-year period showed that of 120 cases aged 60 years or more, 73 (60.8%) were in women and 47 (39.2%) were in men. Now do the proportions of men and women in this sample differ from the proportions of all the other men and women aged over 60 years admitted for surgery over the same period? A random sample of 640 patients admitted for surgery over the period shows that 363 (56.7%) were women and 277 (43.3%) were men. The percentage of women in the burns sample differs from the percentage of women in the general surgical sample by 60.8 - 56.7 = 4.1%. Is this difference significant?

Calculation

The standard error of the difference for two samples 1 and 2 is calculated by:

$$\text{SE diff\%} = \sqrt{\frac{p_1 \times (100 - p_1)}{n^1}} + \sqrt{\frac{p_2 \times (100 - p_2)}{n^2}}$$

Where $n_1$ and $n_2$ are the number in each sample, $p_1$ is one percentage and $100 - p^1$ is the other percentage in sample 1; $p_2$ and $n_2$ represent similar values in sample 2.
So for our example:

$$\text{SE diff \%} = \sqrt{\frac{60.8 \times 39.2}{120}} + \sqrt{\frac{56.7 \times 43.3}{640}} = 4.87$$

The difference between the percentage of women (and men) in the two samples was 4.1%. To find the probability attached to this difference we calculate 95% confidence limits = 2 x the standard error of the difference (see page 202).
The difference between the rates (4.1%) is less than the 95% confidence limits (4.87 x 2 = 9.74). Therefore the difference between the percentages in the two examples could have been due to chance alone.

*Exercise 7*

*Examination of 751 people from village A revealed microfilariae in 310 (41%); similar examinations on 849 people from village B revealed microfilariae in 237 (28%). Is the rate of infection significantly higher in village B?*

Answer 6

| | No. | Mean haemoglobin (g/dl) | Standard deviation |
|---|---|---|---|
| A | 62 | 12.2 | 1.8 |
| B | 35 | 10.9 | 2.1 |

$$\text{SE difference} = \sqrt{\frac{1.8^2}{62} + \frac{2.1^2}{35}}$$

$$= \sqrt{0.052 + 0.126}$$

$$= \sqrt{0.178} = 0.42$$

Therefore, the standard error of the difference between the two means is 0.42 g/dl.
Difference between means = 1.3 g/dl
To calculate probability we divide the difference between means by the SE diff. $= \frac{1.3}{0.42} = 3.095$

By reference to Figure 6 we can see that this value is more than 2.576 which corresponds to a probability of 0.01. Therefore $P < 0.01$

Answer 7

$$\text{SE diff \%} = \sqrt{\frac{41 \times 59}{751}} + \sqrt{\frac{28 \times 72}{849}}$$

$$= \sqrt{3.22 + 2.37}$$

$$= \sqrt{5.59} = 2.36$$

The difference between the positivity rates in the two villages was 41-28 = 13%
To find the probability, we calculate the 95% confidence limits = 2 x 2.36 = 4.72. The difference between the rates is greater than 4.72. Therefore, the positivity rate in village A is significantly higher than that in village B ($P < 0.05$).

## 7.6 CORRELATION

Scatter diagrams

In Unit 5 we considered the use of scatter diagrams for describing the relationship between two variables. Usually in such circumstances, we think of one variable being influenced by the other. It has become conventional to denote the dependent variable, i.e. the one being influenced, by Y and the independent variable by X.

Although scatter diagrams are useful for gaining a visual impression of the relationship, a more quantitative description is often needed. Two kinds of statistical techniques are used to specify further the relationship between X and Y:

- correlation;
- regression.

### CORRELATION COEFFICIENT

Definition

The correlation coefficient, usually denoted by *r*, is an index of the extent to which two variables are associated.

The correlation coefficient has values between +1.0 and -1.0, depending on:

- The strength of the association.
- Whether a positive change in X produces a positive or a negative change in Y.

A correlation coefficient of zero indicates that the two variables are not related.

Study advice

Reference 4 describes and illustrates correlation (Figure 18.1) and also explains how to calculate the correlation coefficient (pages 747-748). Try Exercise 19 on page 748 if you want to practice the calculation.

Figure 7 gives a general guide to interpreting the magnitude of the correlation coefficient.

Figure 7 Interpretation of the correlation coefficient

| Absolute value of *r* | Degree of association |
|---|---|
| 0.8 - 1.0 | Strong |
| 0.5 - 0.8 | Moderate |
| 0.2 - 0.5 | Weak |
| 0 - 0.2 | Negligible |

## REGRESSION

Regression equation

Correlation between two variables means that when one of them changes by a certain amount, the other variable also changes by a certain amount. If Y is the dependent variable and X the independent variable, then the regression equation represents how much Y changes with any given change of X. The equation can be used to construct a regression line on a scatter diagram which allows us to predict Y given a specified level of X.

Regression line

Study advice

Reference 4, pages 802-803 describes how to calculate the regression equation and draw a regression line. Try Exercise 20 on page 803 if you would like to practice the method.

---

# UNIT 8: WRITING AND READING AN EPIDEMIOLOGICAL REPORT

# UNIT 8: OBJECTIVES

Study with this unit will enable you to:

- Write a report of an epidemiological study in a clear, concise and logical manner.
- Critically evaluate the credibility and usefulness of an epidemiological report.

In particular, you will be able to:

- List the major components of an epidemiological report.
- Describe the main features of tables and diagrams in relation to the presentation of findings of an epidemiological report.
- Draw accurate diagrams to illustrate given data.
- List ways in which maps may be used in epidemiological studies.
- Describe the major elements that need to be examined when making a critical assessment of an epidemiological report.
- Demonstrate how to deal with each of these elements with reference to a given published paper.

---

# UNIT 8: CONTENTS

## 8.1 INTRODUCTION

Why write a report?

Your epidemiological study will not be complete until you have produced a report. Even if the results are of interest only to yourself you should make a short, handwritten report, if only to place the findings on record. However, it is likely that you will need to communicate your findings to the central health authority or the district committee, so that they are aware of what you have done and what action needs to be taken as a result. You may even decide that you would like to present your findings at a meeting of your medical association or to publish your work for wider circulation.

If you have spent time and effort planning your study, collecting the data and analysing the results, you will not want your work to go unrecorded and unnoticed, even though you may not enjoy the prospect of producing a report. In this unit we will provide clear and simple guidelines which will help you to write a report (section 8.2) and to present the results as tabulations and/or diagrams (section 8.3).

Making a start

The most difficult part of writing up your study is making a start - actually getting down to putting pen to paper rather than just thinking (and possibly even worrying) about it. Once you have made a start you will find that the task becomes easier and more absorbing, and you will begin to feel that you have something to show for all your previous work.

A rough draft

Your first aim should be to produce a rough draft of your report. This will clarify your thoughts and give you new ideas on how best to express your results and conclusions. Subsequent revisions of your report will be much easier to make.

Discussions

When you have completed your first draft, discuss your report with your colleagues, and your supervisor if you have one. They should be able to give you some useful comments and you should listen seriously to what they have to say. They may point out some important omissions or errors which you have not noticed because you are so concerned with the detail that you cannot clearly see the broader issues. They may have some good ideas for ways in which your report could be improved. If you do not agree with their comments, make sure that you can justify what you have written by well-reasoned arguments.

Evaluation

To be useful, a report should not only give the results of a study and the inferences which have been drawn from them, but also enough information about the methods to enable a critical reader to evaluate the validity of the findings and conclusions, or to check the findings by replicating the study. In the final section of this unit (8.4) we shall consider how to evaluate an epidemiological report. This should assist you not only to critically examine your own work, but also to review the reports of others and to evaluate data or facts

upon which you may have to base priority decisions. As a matter of fact, your skills in critical reading of the medical literature will be one of the most important skills for keeping yourself up to date!

---

## 8.2 WRITING A REPORT

### GENERAL CONSIDERATIONS

The style and nature of your report will depend on the nature and purpose of your study and who is likely to read the report, but the following three general points can be made.

Clarity

- The report should be clearly written, in simple language. Use short sentences and paragraphs, avoid jargon words and present the facts and inferences in a logical sequence.

Length

- Keep the report as short as possible. Include only what is needed to communicate the relevant points. It is not easy to condense a report, especially when you have spent so much time gathering the information which goes into it! You may have to write a number of drafts until you are satisfied, especially if you are not used to writing a report of this kind. But remember, if your report is concise and readable (and interesting) it is much more likely that it will be read and understood by others.

Title

- Choose a good title which clearly explains what the report is about. If necessary add a subtitle for extra clarification. For example:
  'The prevalence of onchocerciasis and blindness in Kifali District.'
  'Epidemiology of schistosomiasis in Aboumi District: relationship between *Schistosoma haematobium* infection and water contact pattern in six villages.'

### CONTENT OF A REPORT

Sections of a report

Conventionally a report contains the following sections. Depending on circumstances, some sections may be subdivided or combined.

1 Introduction
2 Materials and methods
3 Results
4 Discussion
5 Summary
6 References

If you are intending to publish your report, check with the requirements of the journal of your choice as the editor may have listed specific instructions concerning the preparation of the text, tables and illustrations.

#### 1 Introduction.

Purpose

This should state the purpose, topic and objectives of the investigation. Usually the introduction also includes some

Background information

background information such as the nature of the local situation which led to the study. If you have consulted any published

literature or previous work relevant to the study, you should mention this in order to explain the background and significance of the study. Your introduction should also state briefly when and where the study was carried out.

## 2 Materials and methods

In this section you should:

Study population
- Give a description of the study population and its characteristics and, if relevant, information about case-finding methods and the selection of samples (sampling frame, sampling procedure, sample size) and controls (see Unit 3, section 3.4).

Coverage/sampling
- State the coverage or response rate and possible reasons for non-response. Give information on the representativeness of the sample and, if study populations are compared, on their comparability (see Unit 3, section 3.4).

Variables
- List the variables and give the operational definitions, such as criteria used to diagnose a disease or to put a person into a particular occupational group (see Unit 3, section 3.5).

Data collection
- Describe the methods used for collecting the information, including questionnaires. Mention the personnel used and any training given to them (to minimise observer variation). Give the results of any measurements made of observer variation (see Unit 4).

Pilot study
- Describe any pilot study carried out to refine the methodology used and any amendments which were made as a result (see Unit 4).

Analysis
- Describe the method of processing the data and mention any statistical tests used (see Units 5, 6 and 7).

The material and methods section should give the reader enough information to judge the validity and reliability of your methodology.

## 3 Results

Tables and diagrams

In the results section the findings are usually presented as tabulations or diagrams whose main features are also described in the text (see section 6.2).

You should present the facts clearly and concisely. If you think the full results of your study would be of use to other investigators, you could provide additional numerical data in tables as an appendix to your report.

Appendices

Statistical tests

If you have used statistical tests then you should include the results, but only if you are sure that the tests were necessary and appropriate. Unfortunately many people include *P* values because they feel it makes their work look more scientific rather than because they add meaning to the results.

Facts only

As far as possible you should only include facts in this section. However, sometimes you may find it difficult to present consecutive sets of facts intelligibly unless you link them with some interpretation.

**4 Discussion**

Interpretation

In this section you should interpret the results described in the previous section. Make sure you do not introduce new findings at this stage. In the discussion you should also draw comparisons with other relevant studies, and indicate the shortcomings and limitations of your own study. Where appropriate you should consider the practical implications of your findings.

Recommendations

List your recommendations. For example, what further studies should be made as a result of your work, or what practical actions need to be taken.

**5 Summary**

Brief but informative

Also known as an abstract, the summary should be brief but informative. It should include a statement of the objectives, how and on what population the study was performed, and the main findings and recommendations. It is often best to put the summary at the beginning of your report. In that way you will be sure that at least the main points of your study are likely to be read.

**6 References**

Citations

If you have referred to the work of others in your report, you should include a list of references at the end. References are usually cited in the text by stating the author's name and year of publication. It does not matter very much how your references are cited as long as you use a standard system. Arrange the references alphabetically by author's name. If you have two (or more) papers by the same author,

Accuracy

put the one with earliest publication date first. Make sure your citation is correct; anyone who reads your report should be able to find the reference for himself by using your citation. Again, consult the 'Instructions to authors' in the journal of your choice if you intend to submit your report for publication.

*Exercise 1*

*Read quickly through the paper by D.A. Newsome, R.C. Milton and G. Frederique entitled 'High prevalence of eye disease in a Haitian locale' (Reference 5). As you read, identify the features of a report which we have just described.*

You will find that this paper is very clearly set out. Notice how the 'Materials and methods' and 'Results' sections are subdivided so that you can quickly identify how the study was done and what the findings were.

**WRITING YOUR REPORT**

Although we have identified the main sections of an epidemiological report you will not necessarily write them in this order. It is a good idea to begin by writing the 'Materials and methods' section. For one reason, if you have followed our previous recommendations in Unit 3, section 3.6, you should already have part of this section, at least in outline, as your study protocol. Also, this section has a definite structure to it which makes the writing easier. You could

use the information on 'Materials and methods' on page 228 as a checklist of what should be included.

Gaining confidence

You will probably then feel confident enough to write the 'Results' section. Make sure you assemble all the necessary tables, graphs and calculations from your analysis before you begin writing. When you have compiled the results section it will lead you naturally on to the 'Discussion'.

It is usual to leave writing the 'Introduction' until you have written the main body of your report. Write the summary last when you know exactly what the main points for inclusion should be.

Not just epidemiology

We have discussed how to write up an epidemiological study. But what we have said does not just apply to epidemiology. These basic principles are true for any scientific paper. In your future career you will be able to use these guidelines to assist you in writing up any medical investigation. They will also help you to assist others to produce a well-written report of their work.

---

## 8.3 PRESENTING THE RESULTS

### GENERAL PRINCIPLES

Tables and diagrams

The findings of an epidemiological study are usually presented as tabulations and diagrams, which are used as evidence and in order to make the facts clearer and more digestible. It is important to choose those tables and diagrams which will enable you to describe your findings concisely.

Simplicity

Each table and diagram should be designed to demonstrate not more than one or two points. It is better to have several simple tables or diagrams than a single complex one.

Self-explanatory

Each table and diagram should be self-explanatory and comprehensible without reference to the accompanying text. Each must have a concise and informative title, indicating the precise source of the data (such as population, place and date). If you are using several tables and diagrams, each should be identified by a number. For example:

'Figure 2. Two-weekly incidence of diarrhoea in Gambian children under five years, April 1974-1977.'

### TABLES

Tabulation is the simplest method of setting out numerical data. We discussed the use of various kinds of tables for data analysis in Unit 5. Now we will briefly consider a few points to remember when producing the final tables to go into your report.

Guidelines

There are no absolute rules about constructing tables, but certain general principles have become accepted as more or less standard.

Subdivisions

- As the eye is used to moving across the page when reading, it is easier to appreciate data laterally rather than vertically. It is therefore preferable to place the most important subdivisions along the top of the table.

Size

- Tables should be as simple as possible. Two or three small tables are better than a single table containing many details or variables. In general, the maximum number of variables that can be read with ease is three.

Labelling

- A table with its heading and captions should be understandable without reference to the text, but reinforcement or further explanation can be given in the text. The title should be clear and concise; each row and each column should be labelled concisely. Any codes, abbreviations or symbols should be explained in detail in a footnote. See Table 1 on page 40 of the paper by Newsome *et al.* (Reference 5).

Layout

- The title is commonly separated from the body of the table by lines or spaces. In small tables, vertical lines separating the columns may not be necessary.

Sources

- If the data are not your own, their source should be acknowledged in a footnote or in the title.

One final point, remember to check the figures in the table. Make sure the totals tally. Make sure the percentages add up to 100% (not 99.8% or 100.1%!). Do not use too many decimal places when expressing percentages, rates etc. One decimal point will be quite accurate enough for most purposes.

**DIAGRAMS**

Uses

Diagrams are frequently used in the presentation of statistical information. They help to bring out certain aspects of the information, such as patterns or trends, which are not immediately obvious from the casual inspection of tabulated data. Diagrams should clarify and should not complicate or mislead. They should explain the tables and not replace them, unless you are sure the reader will not require numerical data.

General purposes

We considered several types of diagram in Unit 5 and we shall consider them again in more detail shortly, giving you guidelines so that you will be able to draw your own diagrams accurately. But first, here are a few general principles.

Title

- Diagrams, like tables, should have a concise and self-explanatory title.

Axes

- In all graphs and in most diagrams such as histograms and bar charts, the axes should be properly defined. The vertical axis of the diagram is known as the y axis or the ordinate; the horizontal axis is known as the x axis or abscissa. Each of the two axes should be clearly labelled with their scales clearly shown. Frequency is usually presented on the y axis and method of classification on the x axis.

Keys

- If more than one variable or relation has to be shown on the diagram, each should be differentiated clearly by the means of legends or keys. For example see Figure 3 on page 42 of the paper by Newsome *et al.* (Reference 5). Remember: the simplest diagrams are the most effective!

Can you remember the different types of diagram and what they are used for? If you are unsure turn back to Unit 5, sections 5.4 and 5.5 and refresh your memory. Look especially at the figures because now we want to make sure that you know how to draw them.

**How to draw a frequency histogram**

Axes

- The x axis gives a continuous scale of the measurement variable; the y axis represents the frequency.

Bars

- A bar or rectangle is drawn for each class of the grouped data.

Width of bars

- The width of the bar is proportional to the class interval used. If all classes have the same interval, the width of each bar in the histogram is the same (as in Unit 5, Figure 13) and the frequency of each class is represented by the height.

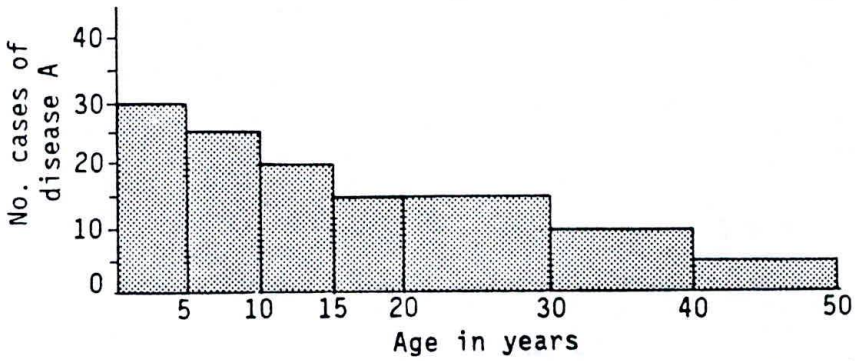
Height of bars

- If unequal class intervals are used, then the height of each bar must be adjusted accordingly. This is illustrated with imaginary data in Figures 1 and 2.

Figure 1 Frequency table showing number of cases of disease A in different age groups

| Age in years | No. cases |
|---|---|
| 0-4 | 30 |
| 5-9 | 25 |
| 10-14 | 20 |
| 15-19 | 15 |
| 20-29 | 30 |
| 30-39 | 20 |
| 40-49 | 10 |

Figure 2 Frequency histogram showing numbers of cases of disease A in different age groups




If the frequency represents two class intervals, then the height of the bar equals half the frequency. If the frequency represents three class intervals, then the height of the bar equals one-third of the frequency, and so on. Note that the class interval 20 to 29 years represents two class intervals (one class interval is five years). Hence the height of the bar is 30/2 = 15. Similarly for class 30-39 years, the height of the bar is 20/2 = 10.

**How to draw a frequency polygon**

- It is usual for the scale on the vertical axis to begin at zero.
- The polygon is constructed from a frequency histogram by connecting the mid-points of the class intervals (see Unit 5, Figures 12 and 13).
- A frequency polygon is derived by first drawing a frequency histogram faintly in pencil. Then join the mid-points on the top of each bar. Finally connect the points and erase the outline of the histogram.

**How to draw a cumulative frequency polygon**

- Obtain a cumulative frequency distribution by adding the frequency of each class in the frequency distribution to the sum

of frequencies of all the preceding classes (see Unit 5, Figure 11).

- For each class plot the cumulative frequency at the end of the class interval for the continuous measurement scale on the x axis.
- Join the points with straight lines to produce the cumulative frequency polygon.

Refer to Unit 5, Figures 12 and 14.

**How to draw a bar chart**

- Frequency may be represented on either the x or the y axis; the other axis represents the discrete variables. Therefore bars may be arranged vertically or horizontally.
- The width of the bars is constant; the length of the bars corresponds to the frequency for that group.
- It is best to arrange the bars in either ascending or descending order of length for ease of reading.
- Bars may be shaded, hatched or coloured in order to emphasize the differences between them.
- When comparisons are made the space between bars in the same group is optional but space between groups is essential (see Unit 5, Figure 8).
- To make a composite bar chart (see Unit 5, Figure 9), each portion of a bar, which represents some sub-division of the group, is directly proportional to the share of that sub-division in the total frequency for the group.

**How to draw a pie diagram**

- First you need to draw a circle of a convenient size. A pie diagram is a circle divided into segments so that the angle of each segment at the centre is proportional to the relative frequency of the variables.
- To convert frequencies to angles, multiply the proportion of each value by 360 (the number of degrees in a circle).
- The convention is to start at the '12 o'clock position' and arrange the segments in the order of their magnitude, largest first, and proceed clockwise around the chart, marking off the angles with a protractor.

Example

The steps are easier to follow by working through an example, using the following data.

| Occupation | Frequency | Sectoral angle |
|---|---|---|
| Unskilled | 120 | A |
| Skilled | 43 | B |
| Professionals | 25 | C |
| Total | 188 | 360° |

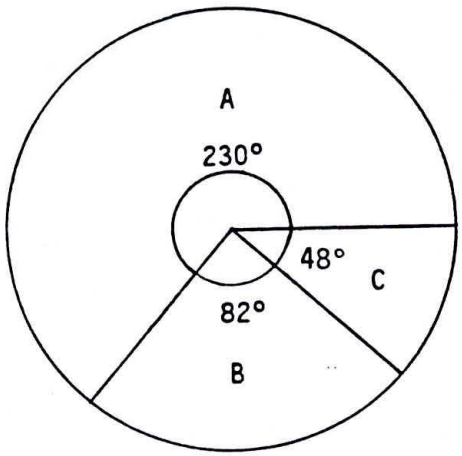
To calculate the sectoral angles:

A = $\frac{120}{188} \times 360° = 230°$

B = $\frac{43}{188} \times 360° = 82°$

C = $\frac{25}{188} \times 360° = 48°$

Figure 3 Pie diagram




**MAPS**

Maps are extremely useful in community health (see Unit 1) and you should already be making good use of the ones on your wall. Simplified maps can also be used for presenting the data in reports.

Maps can be used in the following ways.

Place
- To explain positional data - where the study was carried out. Care must be taken, by the use of insets, to show how small areas relate to the whole country and even continent. For an example look at Figure 1 on page 39 of the paper by Newsome *et al.* (Reference 5). These maps can be used for health service data, to show the distribution of hospitals, health centres, etc.

Physical features
- To illustrate particular physical features, such as rivers or mountains, which may have a special effect on the health problem under study. See page 29 of *Principles of Medicine in Africa* for maps showing the relationship between the diffusion of cholera and physical features.

Population data
- To show population data. This is conventionally done using a dot to represent a specified number of people. A variation of this technique can be used in health service planning: all villages, with a circle to represent their population size, are shaded appropriately if they contain a health facility, and plotted on a map. Areas of poor coverage and their populations can then be seen and new facilities proposed.

Prevalence
- To demonstrate disease prevalence. Variations in degree and coverage of a disease can be very clearly demonstrated using a technique of differential shading.

Comparisons

- To compare data by using a series of maps. For example the distribution of cases of the disease on one map can be compared with the distribution of suspected vector on another, as in the classic study of Burkitt's lymphoma.

Control

- Maps can be used to illustrate the effects of a control programme, for example for *Onchocerca volvulus*, by relating the incidence of infection in children to the sites of larvicidal treatment against the vectors.

Cause

- To study cause. Maps which show associated factors such as rainfall, agricultural areas, educational establishments, communications, or social habits can be useful in elucidating possible cause. See *Principles of Medicine in Africa*, page 32 for maps showing the geographical association between cancer of the oesophagus and alcoholic spirits, and page 92 for a map showing the relationship between endemic goitre and highland areas where soils are deficient in iodine.

If you present your findings on a map, remember to include a key to any symbols or shading you use. Also include an indication of scale.

---

## 8.4 CRITICAL READING OF EPIDEMIOLOGICAL REPORTS

### GENERAL CONSIDERATIONS

In the other units of this module we have discussed the steps involved in any epidemiological study, from the initial identification of the problem, through data collection and analysis and the writing of the report. This section is designed to help you to apply what you have learned to the critical reading of other peoples' reports.

Why be able to read critically?

Why is it important for you to be able to critically evaluate an epidemiological, or indeed any other medical report? In your work you will encounter published (and unpublished) reports of wide-ranging quality and scientific integrity. You may need to evaluate data or facts upon which a policy decision may be based. You may need to distinguish useful from useless (or even harmful) therapy. You may need to decide whether the findings and recommendations from another study can be applied to your own community. In other words, you need to be able to distinguish quickly what is worth reading from what is not, and whether you should change what you do as a result of what you read.

Sound judgement

You therefore need to be able to make sound, independent judgements of the adequacy and reliability of the information, and the validity of the conclusions and the recommendations. You should be able to identify the strengths and weaknesses of any report and to make a balanced, impartial and objective appraisal of its quality. You should not unquestioningly accept all published information, nor should you merely make destructive criticisms. No research is perfect; every study can be constructively criticised.

A critique outline

There is no single and rigid approach to the evaluation of epidemiological or other medical reports. The questions we have used in the critique outline are general, and some of them may not be applicable to every report. For a particular report some questions in the outline will be more relevant than others. Although other important and relevant questions may occur to you, use of the outline will guide you through a systematic and logical appraisal of any report.

Points to consider

Systematic evaluation of a report should consider the following points:

- The objectives, purposes and rationale of the study.
- The methods used for data collection and analysis.
- The data and how they are presented.
- The conclusions and their relevance to the study objectives.
- Any source of error and bias, and how the study could have been improved.

You may also find the outline useful as a checklist in planning your own epidemiological studies.

**OUTLINE FOR EVALUATING AN EPIDEMIOLOGICAL REPORT**

1 **Objectives**

Questions

- Are the objectives of the study clearly stated, vague or unstated? Is the purpose to describe a situation or problem, or to test an hypothesis?
- What is known about the problem? Are the background and magnitude of the problem considered?
- What are the likely strengths/limitations of the study? How could these affect the results/conclusions?

Notes

As a reader, you should have a clear understanding of the objectives of the study. Sometimes you can gain such an understanding just by reading the title of the paper. In many situations the author states his objective or the question to be answered at the beginning of his report. There are cases, however, in which the author does not do so explicitly and you must search out the objective from the conclusions or summary or by reading the entire report. Nevertheless, an initial understanding of the aim of the investigation is central to an effective critical appraisal.

2 **Study design**

Questions

- Is the study design appropriate to answer the study question (e.g. cross-sectional prevalence survey, case control study)? Is the study an experiment, planned observations or an analysis of records?
- What is the population which the investigator intends to study and to which group does he intend to relate his findings? Is this population representative of the disease or the population at risk?
- Is sampling necessary? If so, how is the sample selected? Are there any possible sources of selection bias which will make the sample atypical or non-representative? If so, what provision is made to deal with this bias?
- Is a comparison group(s) necessary and is one used? How is the comparison group selected? Are comparison group subjects and cases comparable for important study variables?

Notes

Careful reading of the 'Materials and methods' section of a paper will allow you to assess the adequacy of the study design. You should first determine the type of study design which is used (see Unit 3, section 3.3) and decide whether this is appropriate for the study problem.

You should then identify the author's target population - the population group to which he intends to refer his findings. Naturally the author will describe the particular group that participated in the study, but it is unlikely that he intends to confine his results solely to the study group. Can you identify the larger group to which he intends to apply the results? If you can, you may then assess whether selection factors or bias could affect the results and hinder conclusions related to the target population.

If the essence of the study is comparison, you need to assess the comparability of the treatment and control groups. If the groups are

not comparable in some important aspect, then the analysis should account for such a discrepancy. Has the author given you sufficient evidence that he is comparing like with like?

3 **Observations**

Questions

- What observations/data collection is planned, on whom, where and when? (Are they consistent?)
- Are these data valid for the objectives of the study?
- Are there clear definitions of the terms used, including diagnostic criteria and measurements made?
- How are data to be collected? Are the data collection techniques (e.g. questionnaires, written records) described, standardised, validated? Are they adequate to obtain accurate, reliable information?

Notes

In assessing the methods of observation you should look out for any inconsistencies in observation or evaluation which could seriously affect the results of the study. In addition to assessing the validity, you should also consider the reliability of the observations, although this may be difficult to assess. Frequently there are some clues in a paper that will at least indicate the author's concern with and awareness of reliability. When a subjective element enters into an assessment, an author will often refer to, and sometimes provide data on, the results of evaluations by independent observers and their degree of agreement. You should be suspicious of results from a study that lacks completely any concern with validity and reliability, especially when the study involves some subjective element either in diagnosis, observation or the assessment of outcome.

4 **Analysis**

Questions

- Is the analysis appropriate for the study objective and the type and level of data collected?
- Are the data worthy of statistical analysis? If so, are the methods of statistical analysis appropriate to the source and nature of the data? Is the analysis correctly performed and interpreted?
- Is there sufficient analysis to determine whether 'significant differences' may in fact be due to faults in methodology or to lack of comparability of the groups in sex or age distribution, in clinical characteristics, or in other relevant variables?

Notes

The important consideration here is whether the design and methods of observation for the study permit the statistical analysis of the results. A major defect in design or some extreme bias in the method of observation can render any assessment of an analysis irrelevant.

5 **Presentation of findings**

Questions

- What results or observations are presented? Are all the findings presented for all subjects? (Were results excluded because of insufficient data, poor response, not statistically significant, etc?)

- Are the results appropriate for the objectives? Are the results correctly interpreted?
- Are the findings presented clearly, objectively and in sufficient detail to enable the reader to judge them for himself?
- Are the findings internally consistent, i.e. do the numbers add up properly? Can the different tables be related to each other, etc.?

Notes

Important findings require proper documentation. For example, if an important result is that males fared better than females, that statement alone is insufficient. The reader is entitled to see the figures or summary statistics so that he may judge for himself that males did indeed do significantly better than females.

It is surprising how often there are numerical inconsistencies contained within papers published in even the most reputable medical journals. Here are some quick checks that you might use for detecting numerical inconsistencies.

- A column should add up to the indicated totals at its foot, and likewise for each row.
- Percentages should add up to 100% if the categories are mutually exclusive.
- Numbers in tables and figures should correspond with those given in the text.
- The totals of various tables should agree, unless the author has specifically indicated deletions to the basic study group, or unless he has noted a restriction to some particular subgroup.

**6 Conclusions**

Questions

- What are the conclusions? Are they justified by the results and analysis?
- Are the conclusions appropriate for the study objectives, study design, population selected for study and the conduct and analysis of the study?
- Is there a discussion of the limitations, errors, bias or problems encountered in the course of the study? Are the findings discussed in relation to these problems? Are alternative explanations of the findings considered?
- Can the conclusions be extrapolated from the study population to the general population?

Notes

The logical and orderly analysis of a paper according to the preceding sections of this outline will lead you to a virtually automatic verdict regarding the author's conclusions. The critical appraisal is essentially completed when you have decided whether the author's conclusions tie in with the objectives as defined in the first section of the outline.

**7 Constructive suggestions**

When you are evaluating a report, do not confine your evaluation to negative criticisms. Consider constructive suggestions as to how the study could have been improved. Assume that you are planning an investigation to answer the questions put in the study. If they have not been clearly put by the authors, formulate them in an

appropriate manner. Suggest a practical design, criteria for observations, and type of analysis that would provide reliable and valid information relevant to the questions under study. However, be reasonable. Make sure the improvements you suggest are realistic. Upon reflection, you may find that there is no feasible alternative study!

*Activity*

*To gain some practice in critical appraisal, read the paper by D.A. Newsome et al. (Reference 5) which we have been studying throughout this unit. Refer to the critique outline as you read. Prepare a written critique of the paper which you can discuss with your supervisor.*

Remember, as much, if not more, can be learned from the critical dissection of a good paper as of one that has obvious major defects.

Acquiring skills

Whenever you read a report or scientific paper, use the critique outline to assist you in evaluating the work which is described and the conclusions which are drawn. With practice, you will be able to remember the questions to ask yourself.

What you can do

Once you have acquired the skills of critical evaluation, you will be able to:

- Distinguish quickly what is worth reading.
- Decide whether you should change your behaviour, practice or policies as a result of what you have read.
- Keep yourself up-to-date with the medical and other literature.

---

## ANSWERS TO THE PRE-TEST

Study advice

> In case you are not entirely satisfied with any one of your answers, we have added a guide to specific further study at the end of each of our answers. For example, *Unit 1* at the end of our answer to Question 1.

*Question 1*

*You wish to study the health of the community in your district, using information that is readily available. List the different kinds of information that you will need.*

- Data on mortality, by cause of death, and by time (e.g. month, season, year), person (e.g. age, sex, occupation) and place (e.g. residence).
- Data on morbidity, by cause, time, person and place.
- Data on population size and structure, i.e. by age, sex, occupation and residence.

*Unit 1*

*Question 2*

*Following on from Question 1, list the places where you might find this information. How reliable is the information that is available to you?*

- Data on mortality: from death certificates and registers. The reliability may be variable depending on how many deaths are actually registered. Also the accuracy of the information about the cause of death will depend upon who completes the certificate.
- Data on morbidity: from hospitals, health centres, disease registers, primary care workers, etc. Again, the reliability depends upon how many sick people actually go along to the hospitals, etc., whether any systematic record is kept of the patients, and whether the diagnoses are accurate.
- Data on population size and structure: this should be available from the national census. The reliability will depend on the accuracy of the census and also any population changes that have occurred since it was carried out due to births, deaths and migration.

*Unit 1*

*Question 3*

*When studying deaths in your district why is it important to calculate rates, in which you relate the number of deaths to the population?*

The number of deaths alone tells us nothing about the risk of dying. Death rates allow you to compare the risk of dying in populations of different size. Epidemiologists always relate the number of deaths (or other events) to the population at risk, to calculate rates. Never compare numbers!

*Unit 1, Section 1.3;*
*Unit 2, Section 2.1*

*Question 4*

*List the different kinds of epidemiological measures you could use to study health and disease in the community where you work. Give the definition of each measure in your list.*

- Death rates. The crude death rate is calculated by relating the total deaths in a year to the total population.
- Incidence rates. Incidence is the number of new cases (or events) of the disease which occur during a specified period.
- Prevalence rates. Prevalence rates measure the number of people in a population who have a disease at a given point in time.
- Case fatality ratio. This is the proportion of persons who die from a disease within a specified time among all those who contract the disease.
- Relative risk. This is the ratio of the incidence rate among the exposed to the incidence rate among those not exposed to the factor.

*Unit 1, Section 1.3;*
*Unit 2, Sections 2.1 and 2.2*

*Question 5*

*You have received reports that there is serious malnutrition among young children in one part of your district. Describe the steps you would take in planning an epidemiological investigation into these reports.*

- State the objectives of the investigation (e.g. to study the prevalence of malnutrition among children in part of your district).
- Define your study population and sampling method (e.g. children under the age of five years in the reported area).
- Select a control population from another part of the district.
- Define the observations to be made (e.g. height, weight, skinfold thickness) and choose a standardised technique for making them.
- Identify potential sources of bias in your investigation (e.g. non-response, observer variation).
- Determine the resources you will need to carry out the investigation (e.g. manpower, training, finance).

*Unit 3*

*Question 6*

*You want to collect data on infant mortality over the last two years in your district. Briefly outline how you will organise the collection of the data.*

- Design appropriate forms, questionnaires, etc. on which to record the data (e.g. birth history form).
- Test the forms in a pilot study.
- Train personnel to collect the data.
- Monitor and evaluate the data collection (e.g. check for systematic error).

*Unit 4*

*Question 7*

*You have been given data on the height (in cm) of 100 boys and 100 girls aged five years in your district.*

*1 How would you present the data on height for each sex diagrammatically?*

*2 How could you summarise the data on height for boys and girls separately?*

1 Frequency distributions for boys and girls separately.

2 Calculate the mean height and standard deviation for boys and girls separately.

*Unit 5, Sections 5.5 and 5.6*

*Question 8*

*Following on from Question 7, you have also been given data on the weight (in kg) of the same 100 boys and 100 girls. For the girls only, how could you examine the relationship between height and weight diagrammatically?*

Plot a scatter diagram of height against weight for the girls only.

*Unit 5, Section 5.7*

*Question 9*

*Following on from Question 8, how could you determine whether, on average, boys weighed more than girls?*

Calculate the mean weight, standard error and 95% confidence limits for boys and girls separately. See whether the confidence limits overlap.

*Unit 7, Section 7.2*

*Question 10*

*Why is it necessary to use standardised death rates when comparing different countries?*

To take into account differences in the age structures of the different populations, since age is an independent determinant of mortality. Whenever you compare populations with different age structures you must age-standardise the rates, to control for the effect of age on mortality (and morbidity).

*Unit 1, Section 1.3*

*Question 11*

*Outline, in not more than 100 words, the major elements that need to be examined when making a critical assessment of an epidemiological report.*

Systematic evaluation of a report should consider the following:

- The objectives, purposes and rationale of the study.
- The methods used for data collection and analysis.
- The data and how they are presented.
- The conclusions and their relevance to the study objectives.
- Any source of error and bias, and how the study could have been improved.

This is just a brief outline of the major elements which need to be considered. Refer to the module for a more detailed critique outline.

*Unit 8, Section 8.3*

---

# INDEX

## Q



## R

# Appendix: Table of random sampling numbers

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 17 | 42 28 | 23 17 | 59 66 | 38 61 | 02 10 | 86 10 | 51 55 | 92 52 | 44 25 |
| 74 49 | 04 49 | 03 04 | 10 33 | 53 70 | 11 54 | 48 63 | 94 60 | 94 49 | 57 38 |
| 94 70 | 49 31 | 38 67 | 23 42 | 29 65 | 40 88 | 78 71 | 37 18 | 48 64 | 06 57 |
| 22 15 | 78 15 | 69 84 | 32 52 | 32 54 | 15 12 | 54 02 | 01 37 | 38 37 | 12 93 |
| 93 29 | 12 18 | 27 30 | 30 55 | 91 87 | 50 57 | 58 51 | 49 36 | 12 53 | 96 40 |
| 45 04 | 77 97 | 36 14 | 99 45 | 52 95 | 69 85 | 03 83 | 51 87 | 85 56 | 22 37 |
| 44 91 | 99 49 | 89 39 | 94 60 | 48 49 | 06 77 | 64 72 | 59 26 | 08 51 | 25 57 |
| 16 23 | 91 02 | 19 96 | 47 59 | 89 65 | 27 84 | 30 92 | 63 37 | 26 24 | 23 66 |
| 04 50 | 65 04 | 65 65 | 82 42 | 70 51 | 55 04 | 61 47 | 88 83 | 99 34 | 82 37 |
| 32 70 | 17 72 | 03 61 | 66 26 | 24 71 | 22 77 | 88 33 | 17 78 | 08 92 | 73 49 |
| 03 64 | 59 07 | 42 95 | 81 39 | 06 41 | 20 81 | 92 34 | 51 90 | 39 08 | 21 42 |
| 62 49 | 00 90 | 67 86 | 93 48 | 31 83 | 19 07 | 67 68 | 49 03 | 27 47 | 52 03 |
| 61 00 | 95 86 | 98 36 | 14 03 | 48 88 | 51 07 | 33 40 | 06 86 | 33 76 | 68 57 |
| 89 03 | 90 49 | 28 74 | 21 04 | 09 96 | 60 45 | 22 03 | 52 80 | 01 79 | 33 81 |
| 01 72 | 33 85 | 52 40 | 60 07 | 06 71 | 89 27 | 14 29 | 55 24 | 85 79 | 31 96 |
| 27 56 | 49 79 | 34 34 | 32 22 | 60 53 | 91 17 | 33 26 | 44 70 | 93 14 | 99 70 |
| 49 05 | 74 48 | 10 55 | 35 25 | 24 28 | 20 22 | 35 66 | 66 34 | 26 35 | 91 23 |
| 49 74 | 37 25 | 97 26 | 33 94 | 42 23 | 01 28 | 59 58 | 92 69 | 03 66 | 73 82 |
| 20 26 | 22 43 | 88 08 | 19 85 | 08 12 | 47 65 | 65 63 | 56 07 | 97 85 | 56 79 |
| 48 87 | 77 96 | 43 39 | 76 93 | 08 79 | 22 18 | 54 55 | 93 75 | 97 26 | 90 77 |
| 08 72 | 87 46 | 75 73 | 00 11 | 27 07 | 05 20 | 30 85 | 22 21 | 04 67 | 19 13 |
| 95 97 | 98 62 | 17 27 | 31 42 | 64 71 | 46 22 | 32 75 | 19 32 | 20 99 | 94 85 |
| 37 99 | 57 31 | 70 40 | 46 55 | 46 12 | 24 32 | 36 74 | 69 20 | 72 10 | 95 93 |
| 05 79 | 58 37 | 85 33 | 75 18 | 88 71 | 23 44 | 54 28 | 00 48 | 96 23 | 66 45 |
| 55 85 | 63 42 | 00 79 | 91 22 | 29 01 | 41 39 | 51 40 | 36 65 | 26 11 | 78 32 |
| 67 28 | 96 25 | 68 36 | 24 72 | 03 85 | 49 24 | 05 69 | 64 86 | 08 19 | 91 21 |
| 85 86 | 94 78 | 32 59 | 51 82 | 86 43 | 73 84 | 45 60 | 89 57 | 06 87 | 08 15 |
| 40 10 | 60 09 | 05 88 | 78 44 | 63 13 | 58 25 | 37 11 | 18 47 | 75 62 | 52 21 |
| 94 55 | 89 48 | 90 80 | 77 80 | 26 89 | 87 44 | 23 74 | 66 20 | 20 19 | 26 52 |
| 11 63 | 77 77 | 23 20 | 33 62 | 62 19 | 29 03 | 94 15 | 56 37 | 14 09 | 47 16 |
| 64 00 | 26 04 | 54 55 | 38 57 | 94 62 | 68 40 | 26 04 | 24 25 | 03 61 | 01 20 |
| 50 94 | 13 23 | 78 41 | 60 58 | 10 60 | 88 46 | 30 21 | 45 98 | 70 96 | 36 89 |
| 66 98 | 37 96 | 44 13 | 45 05 | 34 59 | 75 85 | 48 97 | 27 19 | 17 85 | 48 51 |
| 66 91 | 42 83 | 60 77 | 90 91 | 60 90 | 79 62 | 57 66 | 72 28 | 08 70 | 96 03 |
| 33 58 | 12 18 | 02 07 | 19 40 | 21 29 | 39 45 | 90 42 | 58 84 | 85 43 | 95 67 |
| 52 49 | 40 16 | 72 40 | 73 05 | 50 90 | 02 04 | 98 24 | 05 30 | 27 25 | 20 88 |
| 74 98 | 93 99 | 78 30 | 79 47 | 96 92 | 45 58 | 40 37 | 89 76 | 84 41 | 74 68 |
| 50 26 | 54 30 | 01 88 | 69 57 | 54 45 | 69 88 | 23 21 | 05 69 | 93 44 | 05 32 |
| 49 46 | 61 89 | 33 79 | 96 84 | 28 34 | 19 35 | 28 73 | 39 59 | 56 34 | 97 07 |
| 19 65 | 13 44 | 78 39 | 73 88 | 62 03 | 36 00 | 25 96 | 86 76 | 67 90 | 21 68 |
| 64 17 | 47 67 | 87 59 | 81 40 | 72 61 | 14 00 | 28 28 | 55 86 | 23 38 | 16 15 |
| 18 43 | 97 37 | 68 97 | 56 56 | 57 95 | 01 88 | 11 89 | 48 07 | 42 60 | 11 92 |
| 65 58 | 60 87 | 51 09 | 96 61 | 15 53 | 66 81 | 66 88 | 44 75 | 37 01 | 28 88 |
| 79 90 | 31 00 | 91 14 | 85 65 | 31 75 | 43 15 | 45 93 | 64 78 | 34 53 | 88 02 |
| 07 23 | 00 15 | 59 05 | 16 09 | 94 42 | 20 40 | 63 76 | 65 67 | 34 11 | 94 10 |
| 90 08 | 14 24 | 01 51 | 95 46 | 30 32 | 33 19 | 00 14 | 19 28 | 40 51 | 92 69 |
| 53 82 | 62 02 | 21 82 | 34 13 | 41 03 | 12 85 | 65 30 | 00 97 | 56 30 | 15 48 |
| 98 17 | 26 15 | 04 50 | 76 25 | 20 33 | 54 84 | 39 31 | 23 33 | 59 64 | 96 27 |
| 08 91 | 12 44 | 82 40 | 30 62 | 45 50 | 64 54 | 65 17 | 89 25 | 59 44 | 99 95 |
| 37 21 | 46 77 | 84 87 | 67 39 | 85 54 | 97 37 | 33 41 | 11 74 | 90 50 | 29 62 |

Each digit is an independent sample from a population in which the digits 0 to 9 are equally likely, that is each has a probability of $\frac{1}{10}$.

# Appendix: Table of random sampling numbers

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 16 | 57 04 | 81 71 | 17 46 | 53 29 | 73 46 | 42 73 | 77 63 | 62 58 | 60 59 |
| 98 63 | 89 52 | 77 23 | 61 08 | 63 90 | 80 38 | 42 71 | 85 70 | 04 81 | 05 50 |
| 01 03 | 09 35 | 02 54 | 51 96 | 92 75 | 58 29 | 24 23 | 25 19 | 89 97 | 91 29 |
| 29 07 | 16 34 | 49 22 | 52 96 | 89 34 | 17 11 | 06 91 | 24 38 | 55 06 | 83 59 |
| 72 61 | 80 54 | 70 99 | 24 64 | 11 38 | 83 65 | 27 23 | 40 37 | 84 58 | 48 53 |
| 71 11 | 41 82 | 79 37 | 00 45 | 98 54 | 52 89 | 26 34 | 40 13 | 60 38 | 08 86 |
| 61 05 | 66 18 | 76 82 | 11 18 | 61 90 | 90 63 | 78 57 | 32 06 | 39 95 | 75 94 |
| 81 89 | 42 34 | 00 49 | 97 53 | 33 16 | 26 91 | 57 58 | 42 48 | 51 05 | 48 27 |
| 10 24 | 90 84 | 22 16 | 26 96 | 54 11 | 01 96 | 58 81 | 37 97 | 80 98 | 72 81 |
| 14 28 | 33 43 | 01 32 | 58 39 | 19 54 | 56 57 | 23 58 | 24 87 | 77 36 | 20 97 |
| 35 41 | 17 89 | 87 04 | 28 32 | 13 45 | 59 03 | 91 08 | 69 24 | 84 44 | 42 83 |
| 07 89 | 36 87 | 98 73 | 77 64 | 75 19 | 05 61 | 11 64 | 31 75 | 49 38 | 96 60 |
| 27 59 | 15 58 | 19 68 | 95 47 | 25 69 | 11 90 | 26 19 | 07 40 | 83 59 | 90 95 |
| 95 98 | 45 52 | 27 35 | 86 81 | 16 29 | 37 60 | 39 35 | 05 24 | 49 00 | 29 07 |
| 12 95 | 72 72 | 81 84 | 36 58 | 05 10 | 70 50 | 31 04 | 12 67 | 74 01 | 72 90 |
| 35 23 | 06 68 | 52 50 | 39 55 | 92 28 | 28 89 | 64 87 | 80 00 | 84 53 | 97 97 |
| 86 33 | 95 73 | 80 92 | 26 49 | 54 50 | 41 21 | 06 62 | 73 91 | 35 05 | 21 37 |
| 02 82 | 96 23 | 16 46 | 15 51 | 60 31 | 55 27 | 84 14 | 71 58 | 94 71 | 48 35 |
| 44 46 | 34 96 | 32 68 | 48 22 | 40 17 | 43 25 | 33 31 | 26 26 | 59 34 | 99 00 |
| 08 77 | 07 19 | 94 46 | 17 51 | 03 73 | 99 89 | 28 44 | 16 87 | 56 16 | 56 09 |
| 61 59 | 37 08 | 08 46 | 56 76 | 29 48 | 33 87 | 70 79 | 03 80 | 96 81 | 79 68 |
| 67 70 | 18 01 | 67 19 | 29 49 | 58 67 | 08 56 | 27 24 | 20 70 | 46 31 | 04 32 |
| 23 09 | 08 79 | 18 78 | 00 32 | 86 74 | 78 55 | 55 72 | 58 54 | 76 07 | 53 73 |
| 89 40 | 26 39 | 74 58 | 59 55 | 87 11 | 74 06 | 49 46 | 31 94 | 86 66 | 66 97 |
| 84 95 | 66 42 | 90 74 | 13 71 | 00 71 | 24 41 | 67 62 | 38 92 | 39 26 | 30 29 |
| 52 14 | 49 02 | 19 31 | 28 15 | 51 01 | 19 09 | 97 94 | 52 43 | 22 21 | 17 66 |
| 89 56 | 31 41 | 37 87 | 28 16 | 62 48 | 01 84 | 46 06 | 04 39 | 94 10 | 76 21 |
| 65 94 | 05 93 | 06 68 | 34 72 | 73 17 | 65 34 | 00 65 | 75 78 | 23 97 | 13 04 |
| 13 08 | 15 75 | 02 83 | 48 26 | 53 77 | 62 96 | 56 52 | 28 26 | 12 15 | 75 53 |
| 03 18 | 33 57 | 16 71 | 60 27 | 15 18 | 39 32 | 37 01 | 05 86 | 25 14 | 35 41 |
| 10 04 | 00 95 | 85 04 | 32 80 | 19 01 | 85 03 | 29 29 | 80 04 | 21 52 | 14 76 |
| 23 94 | 97 28 | 60 43 | 42 25 | 26 48 | 48 13 | 34 68 | 39 22 | 74 85 | 03 25 |
| 35 63 | 42 90 | 90 74 | 33 17 | 58 77 | 83 36 | 76 22 | 00 89 | 61 55 | 13 17 |
| 42 86 | 03 36 | 45 33 | 60 77 | 72 92 | 10 76 | 22 55 | 11 00 | 37 60 | 47 73 |
| 67 26 | 92 87 | 09 96 | 85 37 | 82 61 | 39 01 | 70 05 | 12 66 | 17 39 | 99 34 |
| 91 93 | 88 56 | 35 76 | 97 35 | 19 37 | 14 66 | 07 57 | 24 41 | 06 90 | 07 72 |
| 37 14 | 73 35 | 32 01 | 07 94 | 78 28 | 90 33 | 71 56 | 63 77 | 89 24 | 24 28 |
| 07 46 | 50 58 | 08 73 | 42 97 | 20 42 | 64 68 | 48 35 | 04 38 | 28 28 | 36 94 |
| 92 18 | 09 46 | 94 99 | 17 41 | 28 60 | 67 94 | 26 54 | 63 70 | 84 73 | 76 61 |
| 00 49 | 98 43 | 39 67 | 68 40 | 41 31 | 92 28 | 49 57 | 15 55 | 11 81 | 41 89 |
| 08 59 | 41 41 | 33 59 | 43 28 | 14 51 | 02 71 | 24 45 | 41 57 | 22 11 | 79 79 |
| 67 05 | 19 54 | 32 33 | 34 68 | 27 93 | 39 35 | 62 51 | 35 55 | 40 99 | 46 19 |
| 24 99 | 48 06 | 96 41 | 21 25 | 29 03 | 57 71 | 96 49 | 94 74 | 98 90 | 21 52 |
| 65 86 | 27 46 | 70 93 | 27 39 | 64 37 | 01 63 | 21 03 | 43 78 | 18 74 | 77 07 |
| 52 70 | 03 20 | 84 96 | 14 37 | 51 05 | 63 99 | 81 02 | 84 56 | 17 78 | 48 45 |
| 32 88 | 29 93 | 58 21 | 71 05 | 68 58 | 79 08 | 86 37 | 98 76 | 70 45 | 66 23 |
| 54 16 | 39 40 | 98 57 | 02 05 | 65 15 | 73 23 | 51 51 | 75 06 | 38 13 | 51 68 |
| 95 22 | 18 59 | 54 57 | 44 22 | 72 35 | 81 24 | 14 94 | 24 04 | 42 26 | 92 14 |
| 93 10 | 27 94 | 90 45 | 39 33 | 50 26 | 88 46 | 90 57 | 40 47 | 71 63 | 62 59 |
| 19 20 | 85 20 | 15 67 | 78 03 | 32 23 | 50 59 | 24 83 | 64 99 | 18 00 | 78 50 |

Each digit is an independent sample from a population in which the digits 0 to 9 are equally likely, that is each has a probability of $\frac{1}{10}$.

# BREAST-FEEDING CHILDREN IN THE HOUSEHOLD AS A RISK FACTOR FOR CHOLERA IN RURAL BANGLADESH: AN HYPOTHESIS

LEE W. RILEY,[1] STEPHEN H. WATERMAN,[1] A.S.G. FARUQUE,[2] and M.I. HUQ[2]

[1]Division of Bacterial Diseases, Center for Infectious Diseases, Atlanta, Georgia, USA; [2]International Center for Diarrheal Diseases Research, Bangladesh, Dhaka, Bangladesh

Received October 31, 1985
Accepted for publication March 31, 1986

*Abstract. Vibrio cholerae* 01 produces symptomatic and asymptomatic infections. In this study, we investigated a cholera epidemic in northern Bangladesh to specifically search for risks of developing symptomatic infection. A case-control study in six villages found that cases were more likely than controls to have in their family a child who was still breast-feeding and who had been asymptomatic during the epidemic. Among 24 case-control pairs with cholera-like diarrhea as cases, there were 11 discordant for the presence of such a child in the family, in 9 of them, the child was in the case-family (relative risk = 4.5, p = 0.033). Among 13 case-control pairs with laboratory-confirmed cholera as cases, there were 7 discordant for the presence of a breast-feeding child, and in 6 of them, the child was in the case-family (relative risk = 6, p = 0.06). Breast-feeding children in this area are usually kept naked, and defecate onto a cloth pad held against their buttocks by a family member who may be repeatedly exposed to the soiled cloth. Symptomatic infection with *V. cholerae* may depend on exposures to situations that augment the ingested dose of *V. cholerae*, and these findings led us to hypothesize that breast-feeding children, if infected, may play a substantial role (attributable risk = 55%) in facilitating such transmission in rural Bangladesh.

*Key words:* cholera; cholera transmission; breast feeding; *Vibrio cholera*; diarrhea; Bangladesh.

## Introduction

During a cholera outbreak in a community in endemic areas such as Bangladesh, infections with *Vibrio cholerae* 01, the etiologic agent of cholera, may be more widespread than is appreciated since many persons can become asymptomatically infected [1-4]. Although several recent studies in Bangladesh have suggested that most infections during outbreaks of cholera were associated with ingestion of contaminated surface water [1, 5-7], it is not clear from these studies what factors influence development of symptomatic instead of asymptomatic infections. It is possible that clinically-apparent disease may depend more on host factors such as immunity, gastric acidity, or conditions in the community, or family behavioral practices that affect the ingested dose rather than direct exposure to the original source of *V. cholerae* such as water. During the 1983 post-monsoon cholera season, we investigated an El Tor cholera epidemic in northern Bangladesh specifically to search for some risk factors associated with illness to determine what proportion of symptomatic cases could be

attributed to such factors. The investigation led us to hypothesize that breast-feeding children may be an important factor in increasing the ingested dose of *V. cholerae* to a level sufficient to cause symptomatic infection in family members.

## Methods

The investigation was conducted in one administrative district in northern Bangladesh. We began an investigation in this area because in late September to early October 1983, the highest attack rate of diarrhea and deaths caused by diarrhea in Bangladesh, including confirmed cholera, were reported from this area.

To find diarrhea cases, we reviewed local health department records and hospital admission and outpatient clinic logs for cases of diarrhea and diarrheal deaths from the beginning of September 1983. Emergency surveillance had been in effect in this district after September 17, so daily lists of diarrhea cases and deaths in the community were available for review. Since most of the persons with diarrhea did not seek hospital care, the field workers actively sought them by daily visits to every house in their assigned villages.

We used the WHO definition of diarrhea and defined as a case of cholera an illness with watery diarrhea and one or both of the following laboratory findings: stool culture that yielded *V. cholerae* 01, and paired sera that showed a 4-fold or greater rise in vibriocidal antibody titer. We defined as a case of cholera-like diarrhea an illness with watery diarrhea, lasting less than 10 days, and occurring in the same 2-week period and the same village that had cases of laboratory-confirmed cholera. Using these case definitions, we conducted a case-control study by selecting as controls persons without diarrhea in the 2 weeks before interview, matched by age with cholera-like cases from another family in the same cluster of houses surrounding a courtyard (bari). A family was defined as a group of persons eating out of the same cooking pot. Only cases with onset of diarrhea in the 2 weeks before interview were included in the case-control study; if there were multiple cholera-like cases in a family, only the one with the earliest onset of diarrhea (index case) was included. The cases and controls were selected from the village that reported the highest diarrhea attack rate in the district and from five nearby villages. All persons were interviewed by Bengali-speaking physicians and medically-trained technicians.

When our preliminary findings showed that the attack rate was significantly higher among female than among male index cases, we decided to examine if certain behavioral characteristics of female index cases predisposed them to symptomatic infections. Hence, during our second visit to obtain convalescent serum for vibriocidal antibody measurement, we interviewed the same case-control pairs to determine if characteristics such as presence of a breast-feeding child in the family, or certain feeding practices (e.g., feeding the child milk mixed with water) were associated with an index case of cholera or cholera-like diarrhea in the family. In this analysis, we excluded pairs that had a breast-feeding child as a case.

Stool specimens were collected with tellurite-impregnated rectal swabs and plated directly on taurocholate-tellurite-gelatin agar (TTGA) immediately or within a few hours after collection. Suspicious colonies on TTGA were confirmed in the field by testing for agglutination in polyvalent *V. cholerae* antiserum [8]. The specimens were tested further at the International Center for Diarrheal Diseases Research, Bangladesh (ICDDR,B) [9]. Paired fingerstick blood samples (0.05 ml each) were obtained 10 to 14 days apart, diluted 1:10 in normale saline in the field, kept on ice until centrifuged, and then assayed for vibriocidal antibody titers by the method of Benenson [10].

We used a matched-pair analysis (binomial test) to compare cholera cases and cholera-like cases with their respective controls. We determined the best estimate of relative risk (RR), and calculated the 95% one-sided confidence limit from the tail probabilities of the binomial distribution [11]. To estimate the proportion of cases associated with a risk factor, we calculated the attributable risk percentage using the formula of Cole [12].

## Results

Between October 1 and November 6, 619 diarrhea cases with 27 deaths caused by diarrhea were reported from the study district with an overall diarrhea attack rate of 2.2 per 1000 population. Saduadamarhat village with 73 cases had the highest attack rate (34.3 per 1000) among about 150 villages in this district (*figure 1*). The ages of the

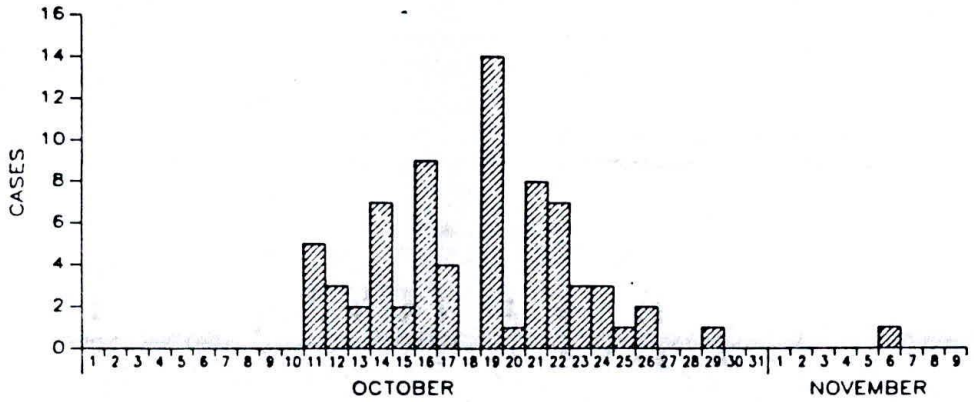



**Figure 1.** Cases of diarrhea recorded in the local health department log book, by date of onset, Saduadamarhat Village, Bangladesh, October-November, 1983.

73 case-patients ranged from 9 months to 70 years with median age of 15 years; 45 (62%) were females. The attack rate for females was 42.1 per 1000 and for males was 25.9 per 1000 ($p < 0.05$, chi square). Cases were distributed throughout the village, often with only one case in each cluster of houses, and were found on both sides of a river that bisected the village.

We identified and interviewed 51 persons in Saduadamarhat village (26) and 5 nearby villages (25) who had cholera-like diarrhea with onset of illness during the peak period of the outbreak (between October 16 and 26). Rectal swab culture and/or paired sera were obtained from 39 (76%) of the patients with cholera-like illness, and the results are shown on *table 1*. Cholera was confirmed in all villages from which cholera-like diarrhea had been reported. Rise in vibriocidal antibody was associated with occurrence of disease; 5 of 13 case-control pairs were discordant for 4-fold or greater rise in vibriocidal antibody titers and in all 5 the rise was in the cases ($p = 0.0312$), binomial test, 1-tailed). Thirteen persons had *V. cholerae* isolated from stool; 2 isolates were serotype Inaba and all others were serotype Ogawa.

We succeeded in identifying age-matched controls for 36 (71%) of the 51 cases. Among these 36 case-control pairs, 20 were from Saduadamarhat village and 13 were

**Table 1.** Results of laboratory tests for *V. cholerae* 01 infections in persons from 6 villages with cholera-like illness, Bangladesh

| *Village* | *Proportion of positive tests among those tested* | | |
|---|---|---|---|
| | *4-fold or greater rise in Vibriocidal antibody titer* | *V. cholerae 01 in stool* | *Either test* |
| Saduadamarhat | 5/16 | 4/11 | 6/21 |
| Tabakpur | 4/ 4 | 4/ 6 | 5/ 6 |
| Bozra | 2/ 2 | 0/ 3 | 2/ 4 |
| Hayat Khan | -/ 0 | 2/ 3 | 2/ 3 |
| Others (2 villages) | 3/ 4 | 3/ 5 | 4/ 5 |
| Total | 14/26 (54%) | 13/28 (46%) | 19/39 (49%) |

**Table 2.** Distribution of case-control pairs by presence in the family of a breast-feeding child, October-November 1983

| *Study group* | *Breast-feeding child in case's family* | *Breast-feeding child in control's family* | | | |
|---|---|---|---|---|---|
| | | *Yes* | *No* | *Total pairs* | *p-value** |
| Proven cholera cases | Yes | 3 | 6 | 9 | 0.061 |
| | No | 1 | 3 | 4 | |
| Saduadamarhat Village cases | Yes | 6 | 7 | 13 | 0.035 |
| | No | 1 | 2 | 3 | |
| Cases in all 6 Villages | Yes | 8 | 9 | 17 | 0.033 |
| | No | 2 | 5 | 7 | |

*binomial test, 1-tail.

pairs that included laboratory-confirmed cholera cases. In the first part of the study, we found no significant differences between cases and controls in their exposures to different sources of water, duration of bathing, frequency of swimming, type of foods eaten, or frequency of contact with a non-family member who had diarrhea in the weeks before date of onset of illness in the case-patient.

We did find that cases were more likely than controls to have in their family an infant or a child who was still breast-feeding and who had been asymptomatic during the epidemic (*table 2*). In all 6 villages, 11 of 24 pairs were discordant for the presence of such a child in the family, and in 9, the child was in the case-family (RR = 4.5, 95% one-sided lower confidence limit = 1.13). In Saduadamarhat village, the relative risk for 8 discordant pairs with cholera-like diarrhea as cases was 7 (95% lower confidence limit = 1.13) and the relative risk for 7 discordant pairs with laboratory-confirmed cholera as cases was 6 (95% lower confidence limit = 0.92). In both case and control-families with such children, similar feeding practices were observed.

To determine if family size of the case-patient was associated with presence of a breast-feeding child in the family, we compared case-families with and without such a child; we did not have family size information for control families. The mean number of members in case-families that had a breast-feeding child was 6.0 and that of families that did not was 5.0. ($p > 0.1$, Wilcoxon Rank Sum test).

The proportion of cases attributable to having a breast-feeding child in the family was 58% for all proven cholera cases, 70% for cholera-like diarrhea cases in Saduadamarhat village, and 55% for cholera-like diarrhea cases in all 6 villages.

## Discussion

In endemic areas, the rate of *V. cholerae* infection in the community during a cholera outbreak may be higher than appreciated [1-4], and the percentage of those infected who develop illness may depend largely on factors such as infectious dose or a variety of host factors or both. This study specifically searched for risk factors for illness, that is, symptomatic infection, and found that illness was associated with presence of a breast-feeding child in the case-family. One possible explanation for this finding is that breast-feeding children who may be protected immunologically may become additional sources of infection for some members of the family who are in more intimate contact with these children. It has been suggested that breast-feeding children are

protected from symptomatic infection by antibody in breast milk, but are not protected from colonization with *V. cholerae* 01 [13]. Thus, children could excrete *V. cholerae* 01 without having diarrhea, and a family member caring for the child could become infected from contact with stool from the child. In rural Bangladesh, small children are usually kept naked and defecate onto a cloth pad held against their buttocks by a family member (usually mothers or older female siblings); sometimes the soiled cloth is folded and used several more times before it is washed. *V. cholerae* has been shown to persist up to 7 days on absorbent material such as cotton [14]. In this way, a family member's hands could repeatedly come into contact with this fecally-contaminated cloth pad.

It is not clear how small children are infected since their diet is limited, but it is possible that during cholera seasons, these children, along with many other persons, ingest the organism in water. The number of organisms ingested may be too small to cause disease in most persons, especially older persons or mothers with partial immunity. However, mothers or older siblings of infected children who repeatedly come into direct contact with the children's feces may ingest larger doses than others and become symptomatic. In the 1950's, it was argued that the presence of children in the household increased the attack rate among women of poliomyelitis (disease transmitted by fecal-oral route [15]. It should be noted that if asymptomatic children are potential risks for others in the family, symptomatic children excreting a higher number of organisms would be even more of a risk. However, in this study, symptomatic children would have been counted as cases, so they would not have been detected as a risk factor. There may be other ways to explain our finding. Mothers who are breast-feeding their children may be immunologically less protected than men of the same age or non-lactating mothers. However, previous studies in Bangladesh have shown no differences in vibriocidal antibody levels (which have been shown to correlate with protection) by sex [4], and a more recent study showed no differences in vibriocidal antibody levels or antitoxic antibody levels between lactating and non-lactating mothers [16].

Another explanation may be that the presence of a breast-feeding child indicates larger family size, which may increase the amount of water used, and hence increased exposures to a potentially-contaminated source. We found that the average family size of the case-patients with a breast-feeding child was not significantly different from that of case-families without such children.

There may be other explanations for our finding, but if our hypothesis is correct, a large proportion of the cholera-like diarrhea cases in the villages we studied (70% in 1, and 55% overall) could be explained by infection, directly or indirectly from presumably asymptomatically infected breast-feeding children. The lack of association between cholera-like diarrhea and specific water or food sources does not rule out the possibility that they were the original sources of *V. cholerae* 01, but even if they were, for clinical and disease control purposes, a factor that influences symptomatic infection, such as that found in this study, may be more important. We present our results as one hypothesis that needs to be tested by a more detailed study.

**We thank Mr. O.G. Siddiqui, Rezaur Rahman, and Shamsul Huda for their microbiologic support, Drs. Hadibur Rahman, Nezam Ahmed, Shahidul Islam, and Haroon-ur-Rashid for their epidemiologic and clinical assistance, Drs. W.B. Greenough, K.M.S. Aziz, M. Huq, and J.D. Clemens, International Center for Diarrheal Diseases Research, Bangladesh and Drs. P.A. Blake, R.I. Glass, J.R. Harris, and R.A. Feldman of the Centers for Disease Control, USA, for their advice and criticism, and the members of the Bangladesh Ministry of Health, District Civil Surgeons, Upazila Health Officers, and most of all, the field workers of Ulipur Upazila who made this study possible.**

Correspondence to: Lee W. Riley, M.D., CID:DBD:EDB: 1-5428, Centers for Disease Control, Atlanta, Georgia 30333.

## References

1. Hughes JM, Boyce JM, Levine RJ, Khan M, Aziz KMA, Huq MI, Curlin CT. Epidemiology of El Tor cholera in rural Bangladesh: importance of surface water in transmission. Bull WHO 1982; 60: 395–404.
2. Woodward WE, Mosley WH. The spectrum of cholera in rural Bangladesh. II. Comparison of El Tor, Ogawa, and Classical Inaba Infections. Am J Epidemiol 1972; 96: 342–51.
3. McCormack WM, Islam S, Fahimmuddin M, Mosley WH. A community study of inapparent cholera infections. Am J Epidemiol 1969; 89: 658–64.
4. Mosley WH, Benenson AS, Barui R. A serological survey for cholera antibodies in rural East Pakistan. I. The distribution of antibody in the control population of a cholera-vaccine field-trial area and the relation of antibody titre to the pattern of endemic cholera. Bull WHO 1968; 38: 327–34.
5. Mosley WH, Khan M. Cholera epidemiology – some environmental aspects. Progress in Water Technology 1979; 11: 309–16.
6. Levine RJ, Khan MR, D'Souza S, Nalin DR. Cholera transmission near a cholera hospital. Lancet 1976; 2: 84–6.
7. Spira WM, Khan MU, Saeed YA, Sattar MA. Microbiological surveillance of intra-neighbourhood El Tor cholera transmission in Bangladesh. Bull WHO 1980; 58: 731–40.
8. Monsur KA. Bacteriologic diagnosis of cholera under field conditions. Bull WHO J963; 28: 387–9.
9. Sommer A, Woodward W. The influence of protected water supplies on the spread of classical/Inaba and El Tor/Ogawa cholera in rural East Bengal. Lancet 1972; 2: 985–7.
10. Benenson AS, Saad A, Paul A. Serological studies in Cholera. I. Vibrio agglutinin response of cholera patients determined by a microtechnique. Bull WHO 1968; 38: 267–76.
11. Breslow NE, Day NE. Statistical methods in cancer research. Vol. I. The analysis of case-control studies. IARC, Lyons 1980.
12. Cole P, MacMahon B. Attributable risk percent in case-control studies. Brit J Prev Soc Med 1971; 25: 242–4.
13. Glass RI, Svennerholm AM, Stoll BJ, Khan MR, Hossain KMB, Huq IM, Holmgren J. Protection against cholera in breast-fed children by antibodies in breast milk. N Engl J Med 1983; 308: 1389–92.
14. Felsenfeld O. Notes on food, beverages and fomites contaminated with *Vibrio cholerae*. Bull WHO 1965; 33: 725–34.
15. Siegel M, Greenberg M, Bodian J. Presence of children in the household as a factor in the incidence of paralytic poliomyelitis in adults. New Engl J Med 1957; 257: 958–65.
16. Glass RI. Abstracts - 19th Joint Conference on Cholera. US-Japan Cooperative Medical Science Program. Bethesda, Maryland 1983; p. 35.

The Lancet · Saturday 2 November 1974

# INFLUENCE OF PASSIVE SMOKING AND PARENTAL PHLEGM ON PNEUMONIA AND BRONCHITIS IN EARLY CHILDHOOD

J. R. T. COLLEY

*Department of Medical Statistics and Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT*

W. W. HOLLAND R. T. CORKHILL

*Department of Community Medicine, St. Thomas's Hospital Medical School, London SE1 7EH*

*Summary* The incidence of pneumonia and bronchitis has been studied in 2205 infants over the first five years of life. In the same period their parents' smoking habits and respiratory symptoms were recorded annually. The incidence of pneumonia and bronchitis in the first year of life was associated with parents' smoking habits; incidence was lowest where both parents were non-smokers, highest where both smoked, and lay between these two levels where only one parent smoked. Over the age of one year the association was not consistent. When parents' respiratory symptoms were also studied a close association was found with the incidence of pneumonia and bronchitis in the child; this was independent of parents' smoking habits and was an almost consistent finding throughout the first five years of life. In the first year of life exposure to cigarette smoke generated when parents smoked doubled the risk for the infant of an attack of pneumonia or bronchitis.

## Introduction

INFANTS who inhale the tobacco smoke generated when their parents smoke at home may have a greater risk of chest illness than the infants of non-smoking parents. We have studied the influence of parental smoking and respiratory symptoms for effects on the incidence of pneumonia and bronchitis in their children during the first five years of life.

## Methods

The data that form the basis of this paper are part of those collected during a longitudinal study of newborn infants and their families. The study was conducted in Harrow, a borough in north-west London, between 1963 and 1969 and involved all families living in six of the wards of the borough who had an infant born in the period July 1, 1963, to June 30, 1965. A total of 2365 families had newborn infants during this period, and, of these, 2205 (93%) were included in the study. The 6·8% excluded (i.e., 160 families) had either moved away from the area before they could be visited or refused to cooperate in the study (table I). The analysis that follows has been based upon the infants born to these families. After exclusions—for example, multiple births—2149 infants were eligible for study. Over the five years of follow-up losses inevitably occurred from the original population; these were small and are unlikely to have seriously biased the findings in the later years of follow-up.

Health visitors, who received special training, administered a questionary to the parents, when, as part of their routine duties, they visited the infant and mother at home within fourteen days of the delivery. At this visit a number of items were recorded, including birth-weight in pounds to the nearest pound below.

TABLE I—SURVEY POPULATION OVER THE FIVE YEARS OF FOLLOW-UP

| No. of families with newborn infants born July 1, 1963, to June 30, 1965 | | No. of index infants at annual follow-up | | | | | |
|---|---|---|---|---|---|---|---|
| Total | Cooperated in survey | Initial visit | First | Second | Third | Fourth | Fifth |
| 2365 | 2205 | 2149 | 2122 | 2109 | 2096 | 2097 | 2095 |

The health visitor also administered a questionary which included questions on respiratory symptoms and smoking habits. In this paper positive responses to the question "Do you usually bring up any phlegm from your chest first thing in the morning in the winter?" has been used as evidence for parental respiratory disability. To elicit smoking habits the questions were: "Do you smoke?" If answered "yes", the parent was classified as a present smoker. If answered "no" the parent was asked "Have you ever smoked?" If the answer was "yes", then the parent was classified as an ex-smoker. If answered "no" the parent was asked "Have you ever smoked as much as one cigarette a day for as long as a year?" An answer "no" classified parents as non-smokers. The present smokers were also asked "How many cigarettes are you smoking now?" The validity of the answers to these questions has already been established.[1]

The families were followed up annually for the next five years by postal questionaries. Each year parents were asked the following questions. For the infant, "Has he/she had in the past twelve months bronchitis? Pneumonia?" For the parents, "Did you usually bring up any phlegm from your chest first thing in the morning last winter?" Smoking habits were assessed using the question "Do you smoke?" If "yes", "How many are

you smoking now?" The validity of answers to the question on infant bronchitis and pneumonia was assessed by checking, in a sample, the parents' account of such an illness with the family doctor's case-notes. The level of agreement was adequate and corresponded to that obtained in other studies where mothers were asked about their children's past health. The validity of the question on phlegm production in the parents has also been established.[8]

In the tables that follow, parents have been classified according to their smoking habits. Parents who at the initial visit had never smoked, and at the first and subsequent follow-ups had not taken up the habit, were classified at each follow-up as non-smokers. In the same way parents who at the initial visit were present smokers, and at the first and subsequent follow-ups did not give up the habit, were classified on each occasion as present smokers. There remained a further group of parents who had changed their habits. These included parents who at the initial visit were ex-smokers. They had been permanently allocated, irrespective of whether or not they took up smoking again, to the "ex-smokers or changed habits" group. In addition there is a further group of parents who were either non-smokers or smokers at the initial visit but who changed their habits during their follow-up. When this occurred they were reclassified permanently as members of the "ex-smokers or changed habits" group. In this way, for example, parents who were smokers at the initial and first and second follow-up visits would be classified as such at these follow-ups. If on the third follow-up they gave up smoking they would be moved to the "ex-smoker or changed habits" group for that and subsequent follow-up years. This method of classification ensures that at each follow-up year the group of "non-smoking" and "present smoking" parents contains parents with consistent smoking habits. The diminishing numbers at each follow-up in these two groups is a result of parents changing their habits and is balanced by the increasing numbers in the "ex-smokers and changed habits" group. The totals in these tables do not correspond to those in table I. This is accounted for by the exclusion of single-parent families and by absent data.

## Results

The annual incidence per 100 children of pneumonia and bronchitis is given in table II by parents' smoking habit. Parents have been classified into one of four groups: (1) both parents non-smokers; (2) one parent smoker, the other non-smoker; (3) both parents smokers; (4) both parents ex-smokers, or one an ex-smoker, or parents who changed their smoking habits during the study. The incidence of pneumonia and bronchitis in the infant shows a gradient by parents' smoking habit in the first year of life. Incidence is lowest in infants with both parents non-smokers, highest where both parents smoke, and lies between these values where one parent smokes. This is a statistically significant gradient ($P<0.0005$). In subsequent years there is no such clear gradient.

TABLE II—PNEUMONIA AND BRONCHITIS BY PARENTS' SMOKING HABITS

| Year of follow-up | Annual incidence per 100 children (absolute numbers in parentheses) of pneumonia and bronchitis | | | | |
|---|---|---|---|---|---|
| | Both non-smokers | One smoker | Both smokers | Both ex-smokers or one ex-smoker or smoking habit changed | All |
| 1 | 7·8 (372) | 11·4 (552) | 17·6 (478) | 9·2 (675) | 11·5 (2077) |
| 2 | 8·1 (358) | 9·3 (494) | 8·9 (438) | 7·4 (758) | 8·3 (2048) |
| 3 | 7·0 (342) | 10·2 (460) | 9·1 (396) | 8·9 (834) | 8·9 (2032) |
| 4 | 8·4 (323) | 8·3 (408) | 9·0 (357) | 8·4 (882) | 8·5 (1970) |
| 5 | 7·5 (319) | 6·7 (374) | 6·5 (340) | 6·6 (956) | 6·7 (1989) |

In table III parents have been classified both by their smoking habits and by their response to the question "Did you usually bring up any phlegm from your chest first thing in the morning in the last winter?" In all categories except one, the incidence within a smoking category is higher among children where one or both parents have winter morning phlegm than in children whose parents are both free of this symptom. Some of the incidence-rates in the children—in particular those whose parents are both non-smokers and who have winter morning phlegm—are based upon small numbers and therefore may not be wholly reliable. On the other hand, the incidence-rates in children where neither parent has symptoms, whether they smoke or not, are based upon substantial numbers. In them in the first year of life a consistent gradient is seen in the incidence of pneumonia and bronchitis in the children in relation to the parents' smoking habits. The rates are lowest in children of non-smoking parents and highest where

TABLE III—PNEUMONIA AND BRONCHITIS IN THE FIRST FIVE YEARS OF LIFE BY PARENTS' SMOKING HABIT AND MORNING PHLEGM

| Year of follow-up | Annual incidence per 100 children (absolute numbers in parentheses) of pneumonia and bronchitis | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Both non-smokers | | One smoker | | Both smokers | | Both ex-smokers or one ex-smoker or smoking habit changed | | All | |
| | N | O/B | N | O/B | N | O/B | N | O/B | N | O/B |
| 1 | 7·6 (343) | 10·3 (29) | 10·4 (424) | 14·8 (128) | 15·3 (339) | 23·0 (139) | 8·2 (546) | 13·2 (129) | 10·1 (1652) | 16·7 (425) |
| 2 | 8·1 (322) | 8·3 (36) | 7·1 (365) | 15·5 (129) | 8·7 (286) | 9·2 (152) | 6·5 (599) | 10·7 (159) | 7·4 (1572) | 11·3 (476) |
| 3 | 6·9 (305) | 8·1 (37) | 10·5 (353) | 9·4 (107) | 7·9 (242) | 11·0 (154) | 8·2 (661) | 11·6 (173) | 8·4 (1561) | 10·6 (471) |
| 4 | 8·0 (287) | 11·1 (36) | 7·5 (306) | 10·8 (102) | 7·6 (236) | 11·6 (121) | 8·2 (695) | 9·1 (187) | 7·9 (1524) | 10·3 (446) |
| 5 | 6·7 (285) | 14·7 (34) | 5·6 (267) | 9·4 (107) | 3·9 (208) | 10·6 (132) | 6·4 (737) | 7·3 (219) | 5·9 (1497) | 9·1 (492) |

N = neither with winter morning phlegm. O/B = one or both with winter morning phlegm.

TABLE IV—PNEUMONIA AND BRONCHITIS BY NUMBER OF CIGARETTES SMOKED PER DAY BY PARENTS AND WINTER MORNING PHLEGM

| Year of follow-up | Annual incidence per 100 children (absolute numbers in parentheses) of pneumonia and bronchitis | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Both non-smokers | | One or both smokers* of following number of cigarettes per day†: | | | | | |
| | | | 1–14 | | 15–24 | | 25 and over | |
| | N | O/B | N | O/B | N | O/B | N | O/B |
| 1 | 7·6 (343) | 10·3 (29) | 10·4 (269) | 15·1 (53) | 11·1 (171) | 14·5 (76) | 15·2 (323) | 23·2 (138) |
| 2 | 8·1 (322) | 8·3 (36) | 5·2 (231) | 16·4 (55) | 8·6 (151) | 14·5 (62) | 9·7 (269) | 9·8 (164) |
| 3 | 6·9 (305) | 8·1 (37) | 11·2 (206) | 8·6 (58) | 8·2 (146) | 9·5 (42) | 8·6 (243) | 11·2 (161) |
| 4 | 8·0 (287) | 11·1 (36) | 5·5 (163) | 13·3 (45) | 7·4 (136) | 11·5 (52) | 9·1 (243) | 10·3 (126) |
| 5 | 6·7 (285) | 14·7 (34) | 6·3 (144) | 11·4 (44) | 4·4 (113) | 7·6 (53) | 4·1 (218) | 10·6 (142) |

* Excluding parent pairs where one or both are ex-smokers or changed smoking habit.
† Includes tobacco and cigars expressed as cigarette equivalents (see Todd [10]).
N = neither with winter morning phlegm. O/B = one or both with winter morning phlegm.

both parents smoke. In children over the age of a year there is, however, no consistent gradient.

Exposure of the child to cigarette smoke may be more precisely estimated from the total daily cigarette consumption of both parents. In table IV the incidence of pneumonia and bronchitis is given for parent pairs smoking between them 1–14, 15–24, and 25 or more cigarettes per day, by the presence of winter morning phlegm. A clear gradient of increasing incidence is seen in the first year of life that is independent of the presence of winter morning phlegm and is of the same size as that in table III. In the second year and thereafter the pattern is not consistent and thus does not suggest an effect of exposure to tobacco smoke at ages over one year.

The gradients of incidence, particularly those attributable to passive smoking in the first year of life, could result from other factors which are known to influence respiratory disease in infancy—for example, social class and family size. These factors might account for the gradients if children of low social class or of large family size were concentrated in families where the parents smoked or had chest symptoms. That these factors did not explain the observed gradient can be seen in tables V and VI. In table V, the findings for social class III alone are examined. The patterns for pneumonia and bronchitis for all children in the first year of life persist. Similarly, in table VI, where the data are subdivided by the number of siblings in the family, the patterns for pneumonia and bronchitis persist within families of the same size. This makes it unlikely that either social class or family size can be responsible for these patterns of respiratory-disease incidence.

The infants of mothers who smoke in pregnancy are, on average, lighter than those of mothers who do not smoke.[3] As infants of low birth-weight are more likely to suffer respiratory illness than normal-weight infants, it is possible that the gradients in respiratory disease observed in the first year of life, and in particular the effects of passive smoking, may be due, indirectly, to maternal smoking during pregnancy. In this study, birth-weight, as expected, shows a

TABLE V—PNEUMONIA AND BRONCHITIS IN THE FIRST YEAR BY PARENTS' SMOKING HABIT AND WINTER MORNING PHLEGM FOR SOCIAL CLASS III

| Annual incidence per 100 children (absolute numbers in parentheses) of pneumonia and bronchitis | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Both non-smokers | | One smoker | | Both smokers | | Both ex-smokers or one ex-smoker | | All | |
| N | O/B | N | O/B | N | O/B | N | O/B | N | O/B |
| 5·9 (171) | 20·0 (15) | 9·5 (263) | 16·5 (79) | 17·1 (217) | 23·9 (88) | 7·1 (294) | 12·1 (66) | 9·8 (945) | 18·2 (248) |

N = neither with winter morning phlegm. O/B = one or both with winter morning phlegm.

gradient by parents' initial smoking habit, and to a lesser extent by winter morning phlegm. Thus parents who smoke have lighter infants than parents who do not smoke. The gradients in the incidence of pneumonia and bronchitis with parental smoking, and with winter morning phelgm, might therefore be partly attributable to differences in birth-weight. However, within different birth-weight categories the gradients for pneumonia and bronchitis with parents' smoking habits persist. Thus differences in birth-weight cannot account for the higher risk of pneumonia and bronchitis in the first year of life in children exposed to the cigarette smoke generated when their parents smoke at home.

## Discussion

An association between the respiratory symptoms

TABLE VI—PNEUMONIA AND BRONCHITIS IN THE FIRST YEAR BY NUMBER OF SIBLINGS AND BY PARENTS' SMOKING HABIT AND WINTER MORNING PHLEGM

| No. of siblings | Annual incidence per 100 children (absolute numbers in parentheses) of pneumonia and bronchitis | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Both non-smokers | | One smoker | | Both smokers | | Both ex-smokers or one ex-smoker or smoking habit changed | | All | |
| | N | O/B | N | O/B | N | O/B | N | O/B | N | O/B |
| 0 | 3·9 (153) | 14·3 (14) | 5·1 (177) | 6·3 (32) | 13·3 (165) | 12·7 (55) | 5·8 (258) | 6·4 (47) | 6·9 (753) | 9·5 (148) |
| 1 | 8·1 (124) | 0 (7) | 13·0 (146) | 12·0 (50) | 13·6 (103) | 34·1 (44) | 9·6 (178) | 17·5 (40) | 10·9 (551) | 19·9 (141) |
| 2 and more | 15·2 (66) | 12·5 (8) | 15·8 (101) | 23·9 (46) | 22·5 (71) | 25·0 (40) | 11·8 (110) | 16·7 (42) | 15·8 (348) | 21·3 (136) |

N = neither with winter morning phlegm. O/B = one or both with winter morning phlegm.

in parents and in their school-age children was reported by Colley.[4] The present study demonstrates that this association is also found in younger children as early as the first year of life. The nature of this association, as Colley noted, is not clear. He concluded that it was unlikely to be an artefact due, for example, to parents with symptoms over-reporting symptoms in their children. In the present study a sample of parents had their account of respiratory illnesses in their children checked against the doctors' records. The close agreement between these two accounts makes it unlikely that over-reporting in families where parents have symptoms has occurred to any important extent.

The association could be a result of shared genetic susceptibility to respiratory disease between parents and children, to living in the same home environment, and to cross-infection within the family. Twin studies in adults have not been notably successful in assessing the genetic contribution to adult chronic respiratory disease, and no studies have yet been reported where this aspect was investigated in parents and their children. The contribution made by the other factors to this association can, at present, only be guessed at.

Passive smoking by the infant, after differences in birth-weight and parental respiratory symptoms have been allowed for, increases the risk to the infant of pneumonia and bronchitis in the first year of life. When both parents smoke, this risk is almost double that of infants with non-smoking parents. The findings confirm and extend those of Harlap and Davies.[5] These workers did not, however, have information on fathers' smoking habits, nor did they take account of parents' respiratory symptoms.

A picture has thus emerged of a serious risk to infants in the first year of life from exposure to their parents' cigarette smoke. In contrast, between one and five years of age, there does not appear to be any important effect of passive smoking in increasing the risk of pneumonia and bronchitis. Colley,[4] in 6–14-year-olds also found no association between passive smoking and the prevalence of chronic cough.

The estimates of children's exposure to cigarette smoke in this study are crude, being based either on whether parents were smokers or not, or on their total daily cigarette consumption. The smoke exposure of the children may have been overestimated, since parents—in particular the father—will smoke outside the home, or at times when the infant is not present. The effects on the child may thus have resulted from exposure to levels of cigarette smoke less than those suggested by our study.

The evidence from this study, taken with that of Harlap and Davies,[5] provides convincing reasons for warning parents who smoke of the risks this entails for their children both from the direct effect of their cigarette smoke and from the presence of their respiratory symptoms. Attacks of pneumonia and bronchitis, particularly in the first year of life, can still result in infant death despite prompt and vigorous treatment. In those that survive such illnesses and recover clinically, the evidence points to some damage to the respiratory tract as indicated by an increased prevalence of chest symptoms and deficits in ventilatory function found in later childhood.[6–8] The longer-term consequences of such childhood illnesses have been underlined by the findings in a cohort of infants followed to the age of 20.[9] At this age the prevalence of chronic cough, after allowing for current smoking habits, social class of father, and air-pollution exposure, was higher in those with a documented history of a chest illness under the age of 2 years than in those without this history. If, by the age of 20, such long-term effects are found, these could persist into middle and late adult life and contribute to the evolution of chronic respiratory disease.

Opportunities for the prevention of serious respiratory disease in infancy and childhood are few. If parents who smoke give up the habit they can reasonably expect to lose, or at least experience an improvement in, their respiratory symptoms. This might well result in reduction of respiratory illnesses in their children. At the same time the absence of cigarette smoke in the home could be expected to diminish the risk of attacks of pneumonia and bronchitis in their children during the first year of life.

This study was conducted jointly with the Health, Welfare, and Children's Department of the London Borough of Harrow, and we would particularly like to thank the Superintendent Health Visitors and their staff, the Senior Administrative Assistant in the Personnel Health Section and his staff, and others who took part for their help and cooperation in this study. Our thanks go to the fieldworkers from the Department of Community Medicine for the maintenance of the records and for their diligence in carrying out the fieldwork during the five years of follow-up. We are also grateful to the statistical assistants of the department for carrying out the analysis of the data.

This study has been supported, in part, by a grant from the Department of Health and Social Security for which we are very grateful.

REFERENCES

1. Holland, W. W., Kasap, H. S., Colley, J. R. T., Cormack, W. *Br. J. prev. soc. Med.* 1969, 2, 77.
2. Krueger, D. R., Rogot, E., Blackwelder, W. C., Reid, D. D. *J. chron. Dis.* 1970, 23, 411.
3. Butler, N. R., Alberman, E. D. (editors). Perinatal Problems. Edinburgh, 1969.
4. Colley, J. R. T. *Br. med. J.* 1974, ii, 201.
5. Harlap, S., Davies, A. M. *Lancet*, 1974, i, 529.
6. Holland, W. W., Halil, T., Bennett, A. E., Elliott, A. *Br. med. J.* 1969, ii, 205.
7. Colley, J. R. T., Reid, D. D. *ibid.* 1970, ii, 213.
8. Bland, J. M., Holland, W. W., Elliott, A. *Revue Physio-path. resp.* (in the press).
9. Colley, J. R. T., Douglas, J. W. B., Reid, D. D. *Br. med. J.* 1973, iii, 195.
10. Todd, G. F. Statistics of Smoking in the U.K. London, 1972.

---

" The atomic physicists were as clever, as modest, as self-seeking, as mean, as argumentative and just as concerned for humanity as the microbe hunters. The physicists worked to produce a weapon of war, but there is no real evidence that the nationalistic arguments which convinced them that their efforts were right and just were any different from those that so affected Koch and Pasteur half a century earlier; and the intellectual challenge was just as great, and grappling with it just as enjoyable. . . . The physicists' work was widely seen as being culpable because it was applied to the taking of life; the first two atomic bombs did so on a vast and horrifying scale. But it was Pasteur, and not some atomic physicist, who in 1870 said of the Germans, ' I want to see the war prolonged into the depths of winter, so that all those vandals confronting us shall perish of cold and hunger and disease '."—ROBERT REID, *Microbes and Men*; p. 168. London: B.B.C. Publications. 1974. £2.50.

# Statistics at Square One

## XIV—The $\chi^2$ tests

T D V SWINSCOW

*British Medical Journal*, 1976, 2, 462-463

The distribution of a discrete variable in a sample often needs to be compared with the distribution of a discrete variable in another sample.

For example, over a period of two years Dr Gold has classified by socio-economic class the women aged 20-64 admitted to his unit suffering from self-poisoning—sample A. At the same time he has likewise classified the women of similar age admitted to a gastroenterological unit in the same hospital—sample B. He has employed the Registrar General's five socio-economic classes, and generally classified the woman by reference to her father's or husband's occupation. The results are set out in table 14.1.

The problem Dr Gold wants to investigate is whether the distribution of the patients by social class differed in these two units. He therefore erects the null hypothesis that there is no difference between the two distributions. This is what he tests by chi-square ($\chi^2$).

It is important to emphasise here that $\chi^2$ tests may be carried out for this purpose only on the *actual numbers* of occurrences, *not* on percentages, proportions, means of observations, or other derived statistics. (There are some quite different purposes for which the $\chi^2$ distribution is used, but they do not concern us here.)

The $\chi^2$ test is carried out in the following steps:

For each observed number (O) in the table find an "expected" number (E); this procedure is discussed below.

Subtract each expected number from each observed number .. .. .. .. .. $O-E$

Square the difference .. .. .. .. .. $(O-E)^2$

Divide the squares so obtained for each cell of the table by the expected number for that cell .. $(O-E)^2/E$

$\chi^2$ is the sum of $(O-E)^2/E$.

To calculate the expected number for each cell of the table consider the null hypothesis. In this case it is that the numbers in each cell are proportionately the same in sample A as in sample B. We therefore construct a parallel table in which the proportions are exactly the same for sample A as for sample B. This is done in columns (2) and (3) of table 14.2. The proportions are obtained from the totals column in table 14.1 and they are then applied to the totals row. For instance, in table 14.2, col (2), $11{\cdot}80=(22\div289)\times155$; $24{\cdot}67=(46\div289)\times155$; and so on. Likewise in column (3) $10{\cdot}20=(22\div289)\times134$; $21{\cdot}33=(46\div289)\times134$; and so on.

Thus by simple proportions from the totals we find an expected number to match each observed number. The sum of the expected numbers for each sample must equal the sum of the observed numbers for each sample, which is a useful check. We now subtract each expected number from its corresponding observed number. The results are given in columns (4) and (5) of table 14.2. Here two points may be noted. The sum of these differences always equals 0 in each column, and each difference for sample A is matched by the same figure, but with opposite sign, for sample B. Again these are useful checks.

Then the figures in columns (4) and (5) are each squared and divided by the corresponding expected numbers in columns (2) and (3). The results are given in columns (6) and (7) of table 14.2. Finally these results, $(O-E)^2/E$, are added. The sum of them is $\chi^2$.

TABLE 14.1—*Distribution by socio-economic class of patients admitted to self-poisoning (sample A) and gastroenterological (sample B) units*

| Socio-economic class | Samples | | Total |
|---|---|---|---|
| | A | B | |
| I | 17 | 5 | 22 |
| II | 25 | 21 | 46 |
| III | 39 | 34 | 73 |
| IV | 42 | 49 | 91 |
| V | 32 | 25 | 57 |
| Total | 155 | 134 | 289 |

TABLE 14.2—*Calculation of chi-square on figures in table 14.1*

| Class (1) | Expected numbers | | O–E | | $(O-E)^2/E$ | |
|---|---|---|---|---|---|---|
| | A (2) | B (3) | A (4) | B (5) | A (6) | B (7) |
| I | 11·80 | 10·20 | 5·20 | −5·20 | 2·292 | 2·651 |
| II | 24·67 | 21·33 | 0·33 | −0·33 | 0·004 | 0·005 |
| III | 39·15 | 33·85 | −0·15 | 0·15 | 0·001 | 0·001 |
| IV | 48·81 | 42·19 | −6·81 | 6·81 | 0·950 | 1·009 |
| V | 30·57 | 26·43 | 1·43 | −1·43 | 0·067 | 0·077 |
| Total | 155·00 | 134·00 | 0 | 0 | 3·314 | 3·833 |

$\chi^2=3{\cdot}314+3{\cdot}833=7{\cdot}147$. DF = 4. $0{\cdot}50>P>0{\cdot}10$.

A helpful technical procedure in calculating the expected numbers may be noted here. Most electronic calculators allow successive multiplication by a constant multiplier to be carried out by a short-cut of some kind. To calculate the expected numbers a constant multiplier for each sample is obtained by dividing the total of the sample by the grand total for both the samples. In table 14.1 for sample A this is $155\div289=0{\cdot}5363\ldots$. This fraction is then successively multiplied by 22, 46, 73, 91, and 57. For sample B the fraction is $134\div289=0{\cdot}4636\ldots$. This too is successively multiplied by 22, 46, 73, 91, and 57. The results are in table 14.2, columns (2) and (3).

Having obtained a value for $\chi^2=\Sigma\{(O-E)^2/E\}$ we look up in a table of $\chi^2$ distribution the probability attached to it. Just as with the *t* table, we must enter the $\chi^2$ table at a certain number of degrees of freedom. To ascertain these requires some care.

**British Medical Journal**
T D V SWINSCOW, MSC, MB, deputy editor

TABLE 14.3—*Distribution of* $\chi^2$

| DF | Probability | | | | | |
|---|---|---|---|---|---|---|
| | 0·50 | 0·10 | 0·05 | 0·02 | 0·01 | 0·001 |
| 1 | 0·455 | 2·706 | 3·841 | 5·412 | 6·635 | 10·827 |
| 2 | 1·386 | 4·605 | 5·991 | 7·824 | 9·210 | 13·815 |
| 3 | 2·366 | 6·251 | 7·815 | 9·837 | 11·345 | 16·268 |
| 4 | 3·357 | 7·779 | 9·488 | 11·668 | 13·277 | 18·465 |
| 5 | 4·351 | 9·236 | 11·070 | 13·388 | 15·086 | 20·517 |
| 6 | 5·348 | 10·645 | 12·592 | 15·033 | 16·812 | 22·457 |
| 7 | 6·346 | 12·017 | 14·067 | 16·622 | 18·475 | 24·322 |
| 8 | 7·344 | 13·362 | 15·507 | 18·168 | 20·090 | 26·125 |
| 9 | 8·343 | 14·684 | 16·919 | 19·679 | 21·666 | 27·877 |
| 10 | 9·342 | 15·987 | 18·307 | 21·161 | 23·209 | 29·588 |
| 11 | 10·341 | 17·275 | 19·675 | 22·618 | 24·725 | 31·264 |
| 12 | 11·340 | 18·549 | 21·026 | 24·054 | 26·217 | 32·909 |
| 13 | 12·340 | 19·812 | 22·362 | 25·472 | 27·688 | 34·528 |
| 14 | 13·339 | 21·064 | 23·685 | 26·873 | 29·141 | 36·123 |
| 15 | 14·339 | 22·307 | 24·996 | 28·259 | 30·578 | 37·697 |
| 16 | 15·338 | 23·542 | 26·296 | 29·633 | 32·000 | 39·252 |
| 17 | 16·338 | 24·769 | 27·587 | 30·995 | 33·409 | 40·790 |
| 18 | 17·338 | 25·989 | 28·869 | 32·346 | 34·805 | 42·312 |
| 19 | 18·338 | 27·204 | 30·144 | 33·687 | 36·191 | 43·820 |
| 20 | 19·337 | 28·412 | 31·410 | 35·020 | 37·566 | 45·315 |
| 21 | 20·337 | 29·615 | 32·671 | 36·343 | 38·932 | 46·797 |
| 22 | 21·337 | 30·813 | 33·924 | 37·659 | 40·289 | 48·268 |
| 23 | 22·337 | 32·007 | 35·172 | 38·968 | 41·638 | 49·728 |
| 24 | 23·337 | 33·196 | 36·415 | 40·270 | 42·980 | 51·179 |
| 25 | 24·337 | 34·382 | 37·652 | 41·566 | 44·314 | 52·620 |
| 26 | 25·336 | 35·563 | 38·885 | 42·856 | 45·642 | 54·052 |
| 27 | 26·336 | 36·741 | 40·113 | 44·140 | 46·963 | 55·476 |
| 28 | 27·336 | 37·916 | 41·337 | 45·419 | 48·278 | 56·893 |
| 29 | 28·336 | 39·087 | 42·557 | 46·693 | 49·588 | 58·302 |
| 30 | 29·336 | 40·256 | 43·773 | 47·962 | 50·892 | 59·703 |

Table 14.3 is taken by permission of the authors and publishers from table IV of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd, London (previously published by Oliver & Boyd, Edinburgh).

When a comparison is made between one sample and another, as in table 14.1, a simple rule for the degrees of freedom is that they equal (number of columns minus 1) × (number of rows minus 1). For Dr Gold's data in table 14.1 this rule gives $(2-1)\times(5-1)=4$. Another way of looking at this is to ask what is the minimum number of figures that must be supplied in table 14.1, *in addition* to all the totals, to allow us to complete the whole table. Four numbers disposed anyhow in samples A and B provided they are in separate rows will suffice.

An abbreviated version of the $\chi^2$ table is given in table 14.3. Entering this at 4 degrees of freedom and reading along the row we find that Dr Gold's value of $\chi^2$, 7·147, lies between 3·357 and 7·779. The corresponding probability is: $0{\cdot}50>P>0{\cdot}10$. This is well above the conventionally significant level of 0·05, or 5%. So the null hypothesis is *not* disproved. It is therefore quite conceivable that in the distribution of the patients between socio-economic classes the population from which sample A was drawn did not differ significantly from the population from which sample B was drawn.

## Quick method

The above method of calculating $\chi^2$ illustrates the nature of the statistic clearly and is often used in practice. But a quicker method devised by Snedecor and Irwin[1] is particularly suitable for use with electronic calculators.

The data are set out as in table 14.1. Take the left-hand column of figures (sample A) and call each observation a. Their total, which is 155, is then $\Sigma a$.

Let p = the proportion formed when each observation a is divided by the corresponding figure in the Total column. Thus here p in turn equals 17/22, 25/46. . . . 32/57.

Let $\bar{p}$ = the proportion formed when the total of the observations in the left-hand column, $\Sigma a$, is divided by the total of all the observations. Here $\bar{p}=155/289$. Let $\bar{q}=1-\bar{p}$, which is the same as 134/289. Then

$$\chi^2=\frac{\Sigma pa-\bar{p}\Sigma a}{\bar{p}\,\bar{q}}.$$

The procedure with table 14.1 is as follows:

Calculate $\frac{17^2}{22}$ and store in memory

" $\frac{25^2}{46}$ " " " "

" $\frac{39^2}{73}$ " " " "

" $\frac{42^2}{91}$ " " " "

" $\frac{32^2}{57}$ " " " "

" $\frac{155^2}{289}$ and subtract from memory.

Withdraw result from memory on to display screen = 1·776975.

We now have to divide this by $\bar{p}\times\bar{q}$. Here $\bar{p}=\frac{155}{289}$ and $\bar{q}=\frac{134}{289}$. So instead of dividing by these fractions we turn them upside down and multiply by them. Thus without removing 1·776975 from the display screen we carry out the following:

$$1{\cdot}776975\times\frac{289}{155}\times\frac{289}{134}.$$

This gives us $\chi^2=7{\cdot}146$.

The calculation naturally gives the same result if the figures for sample B are used instead of those for sample A.

*Exercise 14.* In a trial of a new drug against a standard drug for the treatment of depression the new drug caused some improvement in 56% of 73 patients and the standard drug some improvement in 41% of 70 patients. The results were assessed in five categories as follows: New treatment: much improved 18, improved 23, unchanged 15, worse 9, much worse 8; standard treatment: much improved 12, improved 17, unchanged 19, worse 13, much worse 9. What is $\chi^2$ on this distribution; how many degrees of freedom are there; what is the value of P? *Answer*: 3·295; 4; $P>0{\cdot}5$.

## Reference

[1] Snedecor, G W, and Irwin, M R, *Iowa State College of Science*, 1933, 8, 75.

*What is Takayashu's syndrome?*

**Takayashu's syndrome is a chronic non-specific obliterative arteritis which classically affects the branches of the aortic arch at their origins; hence the synonym "aortic arch syndrome." It is also well known as pulseless disease because all the arm pulses as well as the carotids may be impalpable. A Japanese ophthalmologist, Takayashu, described it in young women. Neurological symptoms are common and may include transient failure of vision associated with the low pressure in the retinal arteries. Since the original description of a disorder affecting the cranial and upper limb arteries, arteriographic studies have shown that there may be multiple stenoses involving the aorta itself, the renal arteries, or the iliac arteries. The same or indistinguishable disorder has been described from Malaysia as the middle aortic syndrome, and the condition is often called "Oriental arteritis" to cover all types of arterial sites, ages, and sexes affected. It is not confined to young women but it is not arteriosclerotic and is seen more often in the young than in the elderly. It is active for a period and then burns out leaving the residual ischaemic legacies which may cause hypertension, a Leriche syndrome, or hemiplegia. It is fairly common everywhere in the Middle and Far East and occurs in South Africa among the Coloured and Black populations. It is very rare in Caucasians but a similar disorder is occasionally seen. The same clinical syndrome has been encountered several times in White patients with systemic lupus. The origin is unknown, there is no specific treatment, but surgical treatment to relieve ischaemia may be helpful in some cases.**

This low number was a surprise and supported the view that its existence encouraged domiciliary care, but in the absence of information about the total work load this conclusion can only be tentative. Another reason for the lower rate of self-referral in our study may be the different nature of medical practice and public attitudes in a large city and a smaller community.

This investigation supports the contention of Patel[2] that the hospital plays a major part in primary emergency care. In these circumstances the public might benefit from an entirely hospital-based emergency medical service in urban areas. The ETS would then be unnecessary and the ambulance service and police relieved of much responsibility. An experienced doctor would receive all emergency calls (replacing the 999 call) from the public after 5 pm and at weekends. His staff would consist of hospital registrars and local general practitioners serving in rotation. On his assessment either an ambulance would be dispatched with or without a doctor in attendance or a general practitioner would make a home call. This system would ensure prompt attention in an emergency and at the same time prevent unnecessary admissions. It would also be an interesting experiment in hospital-general practice integration; a similar system has operated in The Hague since the second world war.[3]

I thank the junior staff members and medical students who assisted in this survey—in particular, Drs J Drury, A Lochrie, and I Fogelman and Mr J Gooden.

### References

1 *British Medical Journal*, 1973, **3**, 248.
2 Patel, A R, *British Medical Journal*, 1971, **1**, 281.
3 *British Medical Journal*, 1976, **1**, 732.

# *Statistics at Square One*

## XV—The $\chi^2$ tests (continued)

T D V SWINSCOW

*British Medical Journal*, 1976, **2**, 513-514

### Fourfold tables

A special form of the $\chi^2$ test is particularly common in practice and quick to calculate. It is applicable when the results of an investigation can be set out in a so-called "fourfold table" or "2 × 2 contingency table."

For example, Dr White, who had been inquiring into the blood pressures of the printers and sheep farmers in her general practice (Part VIII), believed that their wives should be encouraged to breast-feed their babies. She has records for her practice going back over 10 years in which she has noted whether the mother breast-fed the baby for at least three months or not, and these records show whether the husband was a printer or sheep farmer (or some other occupation less well represented in her practice). The figures from her records are set out in table 15.1. The disparity seems considerable, for, while 28% of the printers' wives breast-fed their babies for three months or more, as many as 45% of the farmers' wives did so. What is its significance?

Again the null hypothesis is set up that there is no difference between printers' wives and farmers' wives in the period for which they breast-fed their babies. The $\chi^2$ test on a fourfold table may be carried out by a formula that provides a short-cut to the conclusion. If $a$, $b$, $c$, and $d$ are the numbers in the cells of the fourfold table as shown,

| | | | Total |
|---|---|---|---|
| | $a$<br>$c$ | $b$<br>$d$ | $a+b$<br>$c+d$ |
| Total | $a+c$ | $b+d$ | $a+b+c+d$ |

**British Medical Journal,**
T D V SWINSCOW, MSC, MB, deputy editor

$\chi^2$ is calculated from the following formula:

$$\frac{(a\,d-b\,c)^2\,(a+b+c+d)}{(a+b)\,(c+d)\,(b+d)\,(a+c)}.$$

With a fourfold table there is 1 degree of freedom in accordance with the rule given last week, (number of columns minus 1) × (number of rows minus 1).

Since many electronic calculators have a capacity limited to eight digits, it is advisable not to do all the multiplication or all the division in one series of operations, lest the number become too big for the display. A suitable method is as follows:

| | |
|---|---|
| Multiply $a$ by $d$ and store in memory | |
| Multiply $b$ by $c$ and subtract from memory | |
| Extract difference from memory to display | $a\,d-b\,c$ |
| Square the difference | $(a\,d-b\,c)^2$ |
| Divide by $a+b$ | $\frac{(a\,d-b\,c)^2}{a+b}$ |
| Divide by $c+d$ | $\frac{(a\,d-b\,c)^2}{(a+b)\,(c+d)}$ |
| Multiply by $a+b+c+d$ | $\frac{(a\,d-b\,c)^2\,(a+b+c+d)}{(a+b)\,(c+d)}$ |
| Divide by $b+d$ | $\frac{(a\,d-b\,c)^2\,(a+b+c+d)}{(a+b)\,(c+d)\,(b+d)}$ |
| Divide by $a+c$ | $\frac{(a\,d-b\,c)^2\,(a+b+c+d)}{(a+b)\,(c+d)\,(b+d)\,(a+c)}$ |

With Dr White's figures we have

$$\frac{\{(36\times25)-(30\times14)\}^2\times105}{66\times39\times55\times50}=3{\cdot}418.$$

Entering the $\chi^2$ table with 1 degree of freedom we read along the row and find that 3·418 lies between 2·706 and 3·841.

Therefore $0{\cdot}1 > P > 0{\cdot}05$. So, despite an apparently considerable difference between the printers' wives and the farmers' wives breast-feeding their babies for three months or more, the probability of its occurring by chance is more than 5%.

It should be emphasised again that the $\chi^2$ test is done on the actual numbers of cases, not on, for example, percentages. But suppose the percentages are tested: do we get the same result?

For example, 28% of printers' wives and 45% of farmers' wives breast-fed their babies for three months or more. The difference is 17%. What is the standard error of this difference? It is calculated by the method set out in Part IX. With Dr White's figures we have

$$\text{SE diff }\% = \sqrt{\frac{28 \cdot 72}{50} + \frac{45 \times 55}{55}} = 9{\cdot}24.$$

The difference divided by its standard error is $17/9{\cdot}24 = 1{\cdot}84$. This just falls short of the 1·96 standard errors at the 5% level of probability. Reference to table 7.1 shows that it lies between 1·645 and 1·96, corresponding to $0{\cdot}1 > P > 0{\cdot}05$, the same as with the $\chi^2$ test.

## Small numbers

Experts differ somewhat on how small the numbers in contingency tables may be for a $\chi^2$ test to yield an acceptable result. The following recommendations by Cochran[1] may be regarded as a sound guide. In fourfold tables a $\chi^2$ test is inappropriate if the total of the table is less than 20, or if the total lies between 20 and 40 and the smallest expected (not observed) value is less than 5; in contingency tables with more than 1 degree of freedom it is inappropriate if more than about one-fifth of the cells have expected values less than 5 or any cell an expected value of less than 1.

When the values in a fourfold table are fairly small a "correction for continuity" devised by Yates[2] should be applied. While there is no precise rule defining the circumstances in which to use Yates's correction, a common practice is to incorporate it into $\chi^2$ calculations on tables with a total of under 100 or with any cell containing a value less than 10. Armitage[3] goes so far as to say that "it is probably wise practice to apply it for almost all $\chi^2$ tests for $2 \times 2$ tables." The $\chi^2$ test on a fourfold table is then modified as follows:

$$\frac{\{(|a\,d - b\,c|) - \frac{1}{2}(a+b+c+d)\}^2\ (a+b+c+d)}{(a+b)\,(c+d)\,(b+d)\,(a+c)}.$$

The vertical bars on either side of $a\,d - b\,c$ mean that the smaller of those two products is taken from the larger. Half the total of the four values is then subtracted from that difference to provide Yates's correction. The effect of the correction is always to reduce the value of $\chi^2$.

Applying it to the figures in table 15.1 gives the following result:

$$\frac{\{(36 \times 25) - (30 \times 14) - (105 \div 2)\}^2 \times 105}{66 \times 39 \times 55 \times 50} = 2{\cdot}711.$$

In this case $\chi^2 = 2{\cdot}711$ falls within the same range of P values as the $\chi^2 = 3{\cdot}418$ we got without Yates's correction, $0{\cdot}1 > P > 0{\cdot}05$, but the P value is closer to 0·1 than it was in the previous calculation. In fourfold tables containing lower frequencies than table 15.1 the reduction in P value by Yates's correction may be of considerable significance.

## References


1 Cochran, W G, *Biometrics*, 1954, **10**, 417.
2 Yates, F, *Journal of the Royal Statistical Society*, Supplement, 1934, **1**, 217.
3 Armitage, P, *Statistical Methods in Medical Research*. Oxford, Blackwell Scientific Publications, 1971.


TABLE 15.1—*Numbers of wives of printers and farmers who breast-fed their babies for less than three months or for three months or more*

| | Breast-fed for | | Total |
|---|---|---|---|
| | Up to 3 months | 3 months or more | |
| Printers' wives | 36 | 14 | 50 |
| Farmers' wives | 30 | 25 | 55 |
| Total | 66 | 39 | 105 |

$\chi^2 = 3{\cdot}418$. DF = 1. $0{\cdot}1 > P > 0{\cdot}05$.

*Exercise 15.* An outbreak of pediculosis capitis is being investigated in a girls' school containing 291 pupils. Of 130 children who live in a nearby housing estate 18 were infested and of 161 who live elsewhere 37 were infested. What is the $\chi^2$ value of the difference, and what is its significance? *Answer*: $\chi^2 = 3{\cdot}916$; $0{\cdot}05 > P > 0{\cdot}02$.

The 55 affected girls were divided into two groups of 29 and 26. The first group received a standard local application and the second group a new local application. The efficacy of each was measured by clearance of the infestation after one application. By this measure the standard application failed in 10 cases and the new application in 5. What is the $\chi^2$ value of the difference (with Yates's correction), and what is its significance? *Answer*: $\chi^2 = 0{\cdot}931$; $0{\cdot}50 > P > 0{\cdot}10$.

---

*Two women patients have swollen lips due, apparently, to swelling of the mucous membrane, which has a speckled appearance almost like a pompholyx. No treatment tried has had any effect. What is a possible diagnosis and treatment?*

I cannot be sure of the diagnosis. Is the speckled appearance due to intraepidermal vesiculation (like pompholyx) or could it be due to prominent sebaceous glands (Fordyce spots) or the white streaks of lichen planus? If acute eczematous changes (intraepidermal vesiculation) are present then contact causes should be sought. Lipsticks, toothpastes, miscellaneous sucked objects, and even foods (for example, oranges and artichokes) could be the cause. Patch testing would help in detecting these. Treatment must be directed at removing the offending cause. Glandular cheilitis is a rare condition which might fit the description. It is due to heterotopic salivary glands. Fluid can be massaged out of the little bubbles. Lymphangioma circumscriptum should also be considered, but it is unlikely that one practitioner will see two patients with this. The deep vesicles are characteristic and resemble frog spawn. Treatment for both these conditions can only be surgical. Persistent swelling of a lip, without an obvious infective cause, should also raise the possibility of granulomatous cheilitis and sarcoidosis. If a dermatologist could see the patients and recognise the swelling unnecessary investigations might be prevented.

*Is there any danger to health in the fruit of plants recently treated with systemic insecticide—for example, broad beans for blackfly or gooseberries for caterpillar?*

Before any insecticide (or herbicide) is introduced, official agreement is reached about its toxicity and the conditions for which it should be used.[1] These conditions and the method of use are included in the manufacturer's literature and, provided they are followed, there is no hazard in eating fruit or vegetables which have been sprayed with these compounds. Although these compounds accumulate within the body with time, current ones do not cause any harm. Some of them are highly toxic, however, when inhaled or ingested directly. Practitioners should then contact the Poisons Information Service for advice.


1 Department of Health and Social Security, *Poisonous Chemicals Used on Farms and Gardens—notes for the guidance of medical practitioners*. London, HMSO, 1969.

to recover. All that one can say is that it is likely that many fail to achieve their full potential in adult life. Recent studies have shown that concentration of care on labour and the first weeks of life, which are the most dangerous times in the life of the baby, has paid dividends. By using continuous monitoring for all women, fetal death during labour can be eliminated. Equally, the provision of intensive care for sick babies immediately after birth has resulted in a dramatic improvement in their prognosis. Recognition of these facts has persuaded countries such as France to introduce comprehensive programmes for perinatal care. It seems a tragedy that a country like Britain that has led the world in perinatal research should now decide to cut back on her maternity services. Surely the health of mothers and babies should be one of our major priorities.

### Better use of resources

Just as there is a need to ensure that advances in the care of mothers and babies are maintained, so there is room for making better use of available resources. The progressive improvement in the socioeconomic status and health of the community has been reflected in a change in obstetric practice. Not only are pregnant women healthier but their obstetric performance has improved—they have shorter labours and fewer complications during pregnancy. This should imply that their need for hospital care is less but while recognising that hospital remains the safest place to have a baby, it is now no longer necessary for a mother to stay there for up to 10 days after the delivery.

Experience from units with a high patient-bed ratio has shown that many women with good home circumstances can leave hospital within a few hours of delivery. Although the system requires adequate domiciliary midwifery service backup, it is not only cost-effective (when obstetric beds cost about £20 a day), but it is also desired by modern women.

### Five minutes as a witness

It would be unrealistic to suggest that, after initial savings achieved by such measures, further economic cutbacks are likely to result in anything but a serious decline in the quality of obstetric care. The question we must ask is whether we are wise, as a nation, to give way to the demands of social and political expediency by investing so completely in a policy for improving the lot of the physically and mentally handicapped, yet apparently neglecting the needs of the next generation. The contribution that the NHS will make to the future welfare of the country depends on the selection of the right order of priorities for the next 10 years. Thus it is reasonable to propose that setting up in-depth review of the maternity services should be within the remit of the Royal Commission.

The DHSS has made a start on this important exercise but their proposals need to be examined more critically and in detail. Maternity and midwifery are examples of services that, in recent years, have made valuable contributions to the health of the country and, as such, do not merit the lowest position on the list of priorities.

# *Statistics at Square One*

## XVIII—Correlation

T D V SWINSCOW

*British Medical Journal*, 1976, **2**, 680-681

When two or more series of observations are made it is often found that the observations in one series vary correspondingly with those in the other. The two may increase in parallel, for instance, or decrease in parallel, or as one goes up the other may go down proportionally. This relationship is called correlation. We would say, for example, that the height of children is on the average correlated with age. Since one increases with the other the correlation is called positive. In contrast there is a negative correlation, on the average, between the age of adults and the speed at which they run 100 metres.

The words "on the average" should be noted. In biology we rarely meet examples of perfect correlation, because the sources of the observations, living organisms and their products, vary from one to another and from time to time. Consequently, when we measure the degree of correlation between two sets of observations, we generally find that part of the relationship consists in a true correlation and part consists in random variation due to a multitude of indeterminate causes.

**British Medical Journal**
T D V SWINSCOW, MSC, MB, deputy editor

### Correlation coefficient

The symbol used to denote the coefficient of correlation, as it is called, is r. The correlation to be discussed here, and to which this coefficient applies, is limited to what is called "straight line" correlation. This means that the relationship between the two variables can be expressed graphically by a straight line. Fortunately this is a common feature of correlation in biology. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from +1 through 0 to −1. Complete correlation between two variables is expressed by 1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Fig 18.1 gives some graphical representations of correlation.

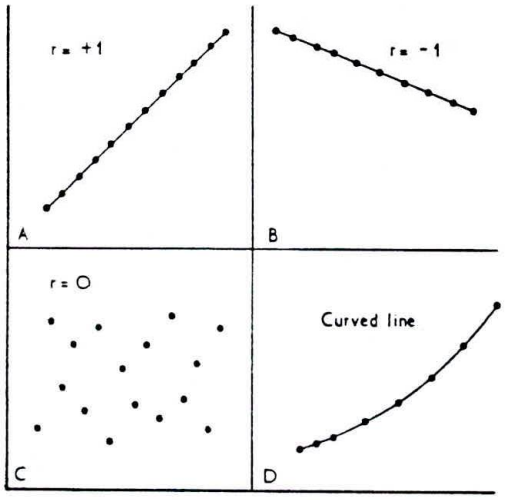



FIG 18.1—Correlation illustrated.

## Scatter diagrams

When an investigator has collected two series of observations and he wants to see whether there is a relation between them, it is best to construct a scatter diagram first. The vertical scale represents one set of measurements and the horizontal scale the other. If one set of observations consists of experimental results and the other consists of a time scale or observed classification of some kind, it is usual to put the experimental results on the vertical axis. These represent what is called the "dependent variable." The "independent variable," such as time or height or some other observed classification, is measured along the horizontal axis, or base line.

The terms "independent" and "dependent" are apt to puzzle the beginner because it is sometimes not clear what is dependent on what. His confusion is a triumph of common sense over misleading terminology, because often enough each variable is dependent on some third variable, which may or may not be mentioned. It is reasonable, for instance, to think of the height of children as dependent on age rather than the converse. But consider a positive correlation reported by Russell *et al*[1] between mean tar yield and nicotine yield of certain brands of cigarette. The nicotine liberated is unlikely to have its origin in the tar: probably both vary in parallel with some other factor or factors in the composition of the cigarettes. The yield of the one does not seem to be "dependent" on the other in the sense that, on the average, the height of a child depends on his age. In such cases it often does not matter which scale is put on which axis of the scatter diagram. However, if the intention is to make inferences about one variable from the other, the observations *from which* the inferences are to be made are usually put on the base line.

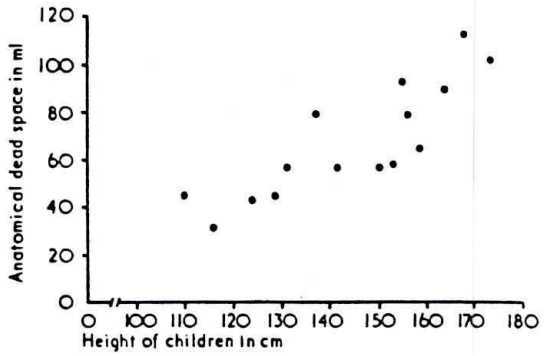



FIG 18.2—Scatter diagram of relation in 15 children between height and pulmonary anatomical dead space.

In practice the dots in a scatter diagram generally lie neither in a single straight line nor equidistant on either side of a central line but in a roughly elliptical area. For example, Kerr[2] has shown that the "anatomical dead space" in the lungs of normal children is positively correlated with age. In other words, the older the child the larger is this space. Dr Green, a paediatric registrar (Part I), has measured the pulmonary anatomical dead space (in ml) and height (in cm) of 15 children, and prepared the scatter diagram shown in fig 18.2. Each dot represents one child, and it is placed at the point corresponding to the measurement of the height (horizontal axis) and the dead space (vertical axis). He now inspects the pattern to see whether it seems likely that the area covered by the dots centres on a straight line or whether a curved line is needed to go through its centre. In this case Dr Green decides that a straight line can adequately describe the general trend of the dots. His next step will therefore be to calculate the correlation coefficient.

## References

[1] Russell, M A H, *et al*, *British Medical Journal*, 1975, **3**, 71.
[2] Kerr, A A, *Thorax*, 1976, **31**, 63.

---

*Do apples contain a natural diuretic?*

I do not know of one and cannot find it in any of the reference books on toxicants naturally occurring in foods. The scientists at a major firm that makes apple juice and at a fruit research station tell me that nothing has ever appeared in print on the diuretic properties of apples and apple juice. The only information that may be relevant is a small newspaper article reporting that an apple diet, used in Switzerland, lowers blood pressure (I have not yet seen this in a scientific journal). Apples have a very low sodium content, and if the patients lost weight this could be an additional explanation.

*Is there any place for ultrasonics in treating Menière's disease?*

Ultrasonic irradiation of the labyrinth for Menière's disease was first introduced by Arslan[1] in 1953 and its value is well established. It is indicated for patients not responding to medical treatment in whom there is useful residual hearing. The operation was originally designed for purely destructive purposes, but it also helps the hydrodynamics of the endolymph system. The surgical equipment required is complicated and requires a high degree of maintenance. Various techniques have been used including irradiation of the lateral semicircular canal, vestibule, and oval window by Arslan[2] and the round window by Kosseff.[3] The results of operation are continually improving and Arslan claims that attacks of vertigo persist in only 10% of patients when in experienced hands and with up-to-date equipment. Facial paralysis which has been an initial complication has been reduced to 0·5%. In the experience of Angell James[4] in Britain, who has irradiated 415 cases of Menière's disease, including many difficult and problem cases, vertigo was relieved in 85%, tinnitus relieved in 70%, and caloric responses abolished in 78%. There had been some further reduction in hearing in 40%, and total loss of hearing had occurred in 2%. In less experienced hands results have unfortunately been less satisfactory.[5]

[1] Arslan, M, in *Proceedings of the Fifth International Congress of Oto-rhino-laryngology, 1953*, ICS No 2, p 429. Amsterdam, Excerpta Medica, 1955.
[2] Arslan, M, *Journal of Laryngology and Otology*, 1970, **84**, 131.
[3] Kosseff, G, Wordsworth, J H, and Dudley, P F, *Archives of Otolaryngology*, 1967, **86**, 535.
[4] Angell James, J, *Archives of Otolaryngology*, 1969, **89**, 95.
[5] Morrison, W W, *Management of Sensorineural Deafness*, p 163. London, Butterworths, 1975.

asked about any case, both for information and to establish why an unusual procedure has been carried out. Most cases are not discussed in detail as they are uncomplicated and not of particular clinical interest. Important complications and misdiagnoses are, however, discussed in considerable depth when the case warrants it. Radiological investigations are also reviewed. A free and open discussion then takes place to discover if and when errors of judgement took place. Deaths are treated in a similar fashion, and particular attention is paid to those cases where the underlying condition was not fatal. Two examples that were freely discussed in this way during the last year were acute appendicitis in a young patient who later developed fatal septic shock and a patient who died from a duodenal fistula. This developed in a chronically malnourished man after simple closure of a large perforated duodenal ulcer.

### Results

In the first year of the review the general surgeons in the Salford Area reported the admission of 5774 patients, including patients admitted as day cases. There were 4440 operations carried out with a reported complication rate of 5·8%. The 215 patients who died during the year included those admitted to the wards for terminal care.

On reviewing the misdiagnoses, we found that very few had serious implications. Thus the confusion of acute gynaecological disorders with acute appendicitis, a commonly recognised diagnostic problem, was reported on 15 occasions. Only rarely were serious misdiagnoses reported.

Initially the audit was somewhat subdued because all surgeons, especially the junior ones, felt that open discussion was critical of both themselves and their colleagues. This inspired a rather defensive reaction to begin with, but as the weeks passed and members of surgical teams realised that the discussion was essentially fair and non-aggressive, the atmosphere changed. Junior surgeons commented that it was useful to hear the more experienced surgeons indicating their own approach to problems such as haematemesis, the management of abdominal injuries, and whether certain investigations were valuable in some crises.

### Discussion

"No clinician should be averse to some form of feedback of performance that permits an external critique."[2] The complex auditing procedures carried out in the United States are not at present applicable in this country, for they demand expensive and extensive administrative machinery and their value has yet to be proved. Nevertheless, the type of examination of surgical practice we have described is possible in both district general and teaching hospitals. It can be carried out only by doctors who understand the problems being discussed, and for this reason it should be open only to the staffs of the units concerned. Occasionally, however, visiting surgeons may be invited to join in the audit and discuss the cases. The attendance of anyone else would undoubtedly inhibit the frank discussions that take place. It is important that those in charge of the auditing procedure must not be aggressive, especially to the junior staff, who may have had to make difficult decisions in less than ideal circumstances. An audit such as we have described undoubtedly has a positive educational role as the experience of the more senior surgeons is pooled during discussion of difficult problems. The main difficulty would appear to be that of boredom and repetition, yet there is usually sufficient variety of material each week to provide a lively clinical discussion on the week's work.

One further advantage is that an audit such as this gives the opportunity for differences in surgical practice to be aired without fear or favour. For example, some surgeons routinely carry out sigmoidoscopy under general anaesthesia—is this necessary? Another example is the treatment of pilonidal sinus—is wide surgical excision, leaving the wound to granulate, justifiable now that minor procedures, such as Lord's operation, are available? Finally, difficult ethical decisions, such as the wisdom and efficacy of operating on very old patients with ruptured aortic aneurysms, may also be freely discussed.

To undertake an audit less frequently than once a week would mean that many cases would be left undiscussed simply because of the volume of work that passes through the units. Regular monitoring of the clinical work load should not, however, detract from or replace normal postgraduate activities such as clinico-pathological conferences to discuss in detail the more interesting cases. With the establishment in a hospital of such a group a way is open for the health authorities to refer statistics derived from the Hospital Activity Analysis for discussion.

We acknowledge the co-operation of our surgical colleagues Mr J R N Curt, Mr A W Hargreaves, Mr G Ingram, and Mr R W Marcuson, and their junior staff, without whose help this venture could not have been started.

### References

1 Slee, U N, *Annals of Internal Medicine*, 1974, **81**, 97.
2 Dudley, H A F, *Proceedings of the Royal Society of Medicine*, 1975, **68**, 634.

# *Statistics at Square One*

## XIX—Correlation (continued)

T D V SWINSCOW

*British Medical Journal*, 1976, **2**, 747-748

### Calculation of correlation coefficient

When making the scatter diagram (Part XVIII, fig 18.2) to show the heights and pulmonary anatomical dead spaces in the 15 children he was studying, Dr Green set out the figures as in cols (1), (2), and (3) of table 19.1. It is helpful to arrange the observations, as he has done, in serial order of the independent variable when one of the two variables is clearly identifiable as independent. The corresponding figures for the dependent variable can then be examined in relation to the increasing series for the indepenent variable. In this way we get the same picture, but in numerical form, as appears in the scatter diagram.

The calculation of the correlation coefficient is as follows, with x representing the values of the independent variable (in this case height) and y representing the values of the dependent variable

**British Medical Journal**
T D V SWINSCOW, MSC, MB, deputy editor

(in this case anatomical dead space). The formula to be used is

$$r=\frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}.$$

The computation is as follows:

| | |
|---|---|
| Note number of observations of x | $n$ |
| Find the sum of the x observations | $\Sigma x$ |
| Find the mean of the x observations | $\bar{x}$ |
| Find the sum of the squares of the x observations | $\Sigma x^2$ |
| Find the square of the sum of the x observations | $(\Sigma x)^2$ |
| Divide this by n | $\frac{(\Sigma x)^2}{n}$ |

Find the sum of the squared differences of the observations from the mean, $\Sigma(x-\bar{x})^2$, from the identity $\Sigma x^2-\frac{(\Sigma x)^2}{n}$.

This procedure is exactly the same as in finding the standard deviation, except that we here stop at finding the sum of the squared differences between the observations and their mean, $\Sigma(x-\bar{x})^2$.

The procedure is then repeated for the y observations, so that from the identity $\Sigma(y-\bar{y})^2=\Sigma y^2-\frac{(\Sigma y)^2}{n}$ we likewise have the sum of the squared differences of the y observations from their mean.

| | |
|---|---|
| Multiply $\Sigma(x-\bar{x})^2$ by $\Sigma(y-\bar{y})^2$ and take the square root | $\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}$ (1) |
| This gives us the denominator of the formula. | |
| To obtain the numerator we proceed as follows: | |
| Multiply each value of x by the corresponding value of y and add these products together | $\Sigma xy$ (2) |
| Multiply the sum of the x observations by the sum of the y observations | $\Sigma x\times\Sigma y$ |
| Divide this product by the number of *pairs* of observations | $(\Sigma x\times\Sigma y)/n$ (3) |
| Subtract (3) from (2) | $\Sigma xy-\frac{\Sigma x\times\Sigma y}{n}$ |
| This is identical to $\Sigma(x-\bar{x})\,(y-\bar{y})$ (4) | |
| Finally, divide (4) by (1) to give the correlation coefficient, r. | |

The calculations on Dr Green's data follow—see also table 19.1, columns (4), (5), and (6). In practice the separate values for $\Sigma x^2$, $\Sigma y^2$, and $\Sigma xy$ are not written down as they are in the table but accumulated in the calculator to form the totals given at the foot of those columns.

$n=15$ $\quad$ $n=15$

$\Sigma x=2\,169$ $\quad$ $\Sigma y=1\,004$

$\bar{x}=144{\cdot}6$ $\quad$ $\bar{y}=66{\cdot}93$

$\Sigma x^2=318\,889$ $\quad$ $\Sigma y^2=75\,030$

$(\Sigma x)^2/n=313\,637{\cdot}4$ $\quad$ $(\Sigma y)^2/n=67\,201{\cdot}07$

$\Sigma(x-\bar{x})^2=$ $\quad$ $\Sigma(y-\bar{y})^2=$

$\Sigma x^2-\frac{(\Sigma x)^2}{n}=5\,251{\cdot}6.$ $\quad$ $\Sigma y^2-\frac{(\Sigma y)^2}{n}=7\,828{\cdot}93.$

$$\Sigma xy=150\,605$$

$$(\Sigma x)\,(\Sigma y)/n=145\,178{\cdot}4$$

$$\Sigma(x-\bar{x})\,(y-\bar{y})=\Sigma xy-\frac{(\Sigma x)\,(\Sigma y)}{n}=5\,426{\cdot}6.$$

$$r=\frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}=\frac{5\,426{\cdot}6}{\sqrt{5\,251{\cdot}6\times7\,828{\cdot}93}}=0{\cdot}846.$$

**The correlation coefficient of 0·846 indicates a strong positive correlation between size of pulmonary anatomical dead space and height of child. But in interpreting correlation it is important to remember the familiar adage, *correlation is not causation.* There may or may not be a causative connection between the two correlated variables. Moreover, if there is a connection it may be indirect, so that their mutual variation has a common cause elsewhere.**

A part of the variation in one of the variables (as measured by its variance) can be thought of as being due to its relationship with the other variable and another part as due to undetermined (often "random") causes. The part due to the dependence of one variable on the other is measured by $r^2$. In Dr Green's investigation $r^2=0{\cdot}716$. So we can say that 72% of the variation in size of the anatomical dead space is accounted for by the height of the child.

TABLE 19.1—*Correlation between height and pulmonary anatomical dead space in 15 children*

| Child number (1) | Height in cm x (2) | Dead space in ml y (3) | Col (2) squared $x^2$ (4) | Col (3) squared $y^2$ (5) | Col (2) × col (3) x y (6) |
|---|---|---|---|---|---|
| 1 | 110 | 44 | 12 100 | 1 936 | 4 840 |
| 2 | 116 | 31 | 13 456 | 961 | 3 596 |
| 3 | 124 | 43 | 15 376 | 1 849 | 5 332 |
| 4 | 129 | 45 | 16 641 | 2 025 | 5 805 |
| 5 | 131 | 56 | 17 161 | 3 136 | 7 336 |
| 6 | 138 | 79 | 19 044 | 6 241 | 10 902 |
| 7 | 142 | 57 | 20 164 | 3 249 | 8 094 |
| 8 | 150 | 56 | 22 500 | 3 136 | 8 400 |
| 9 | 153 | 58 | 23 409 | 3 364 | 8 874 |
| 10 | 155 | 92 | 24 025 | 8 464 | 14 260 |
| 11 | 156 | 78 | 24 336 | 6 084 | 12 168 |
| 12 | 159 | 64 | 25 281 | 4 096 | 10 176 |
| 13 | 164 | 88 | 26 896 | 7 744 | 14 432 |
| 14 | 168 | 112 | 28 224 | 12 544 | 18 816 |
| 15 | 174 | 101 | 30 276 | 10 201 | 17 574 |
| Total | 2 169 | 1 004 | 318 889 | 75 030 | 150 605 |
| Mean | 144·6 | 66·93 | | | |

## Standard error

The correlation coefficient also has a standard error, which for large samples is approximately $\frac{1-r^2}{\sqrt{n}}$.

However, to test the deviation of r from 0, or nil correlation, it is better to use the *t* test in the following calculation:

$$t=r\sqrt{\frac{n-2}{1-r^2}}.$$

The *t* table is entered at $n-2$ degrees of freedom.

For example, the correlation coefficient for Dr Green's figures was 0·846. The number of pairs of observations was 15. Applying the above formula, we have

$$t=0{\cdot}846\times\sqrt{\frac{15-2}{1-0{\cdot}846^2}}=5{\cdot}72.$$

Entering the *t* table at $15-2=13$ degrees of freedom we find that, at $t=5{\cdot}72$, $P<0{\cdot}001$. So the correlation coefficient may be regarded as highly significant.

*Exercise* 19. In a part of Kenya the incidence of kala-azar in homesteads was related to the proximity of termite hills.[1] Presumably this was because the sandfly that transmits the protozoon causing the disease found shelter there. A doctor wanting to know whether this relationship was also true for another part of Kenya visited 16 homesteads in which 40 or more people lived and for each homestead made two measurements: (1) the percentage of people suffering from kala-azar; (2) the mean distance of the five nearest termite hills. The two observations for each homestead were as follows: (1) 21%, 68m; (2) 12%, 103m; (3) 30%, 17m; (4) 8%, 142m; (5) 10%, 88m; (6) 26%, 58m; (7) 42%, 21m; (8) 31%, 33m; (9) 21%, 43m; (10) 15%, 90m; (11) 19%, 32m; (12) 6%, 127m; (13) 18%, 82m; (14) 12%, 70m; (15) 23%, 51m; (16) 34%, 41m. What is the coefficient of correlation between percentage of kala-azar cases and mean distance of termite hills? *Answer:* $r=-0{\cdot}848$.

## Reference

[1] Southgate, B A, and Oriedo, B V E, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1962, **56**, 30.

TREATMENT SUGGESTIONS

*Twenty-four-hour lifeline*—All members of the practice must be aware of the families at risk, even if they are not directly concerned with them. It has been possible to encourage the families to establish links with the whole practice and to use the doctor on duty if a crisis occurs when the primary therapist is absent.

*Therapeutic relationship*—This is provided by the team; the doctor or health visitor acts as the primary therapist and sees the family either at home or in the health centre. In some cases both husband and wife are seen regularly. In others, the mother attends a therapeutic group with an attached playgroup, which has been established recently in the practice. This will be described elsewhere.

*Child care*—The child is seen regularly either at home or in a clinic. Most parents of children at risk have rich fantasies and unrealistic expectations of their child's capabilities and development. These must be gradually and gently brought nearer to reality, which may take weeks or months.

*Practical help*—We have found the most useful form of practical help to be the provision of a playgroup/nursery place. Help with transport, baby-sitting, domestic arrangements, social activities, and, in extreme cases, housing has alleviated family stress and reduced the risks to the child. We try to maintain an honest and realistic approach to these problems to avoid raising false hopes and expectations.

*Referral to other agencies*—Informal discussion with members of the social services is often as important as a formal referral. Even after referral the responsibility for the case is shared. In some cases, psychiatric or paediatric specialist services, or both, are necessary. Each member of the primary health care team has a continuing contribution to make to assessment and management.

OUTCOME

It is too soon to know how much such an approach will reduce the prevalence of actual abuse. We know that two of the 30 at-risk children have suffered minor inflicted injuries—a bruise and a red slap mark on the face. Both of these would have passed unnoticed without the extra attention the families were receiving.

We are confident that all the families have benefited from our intervention, particularly those mothers and children attending the therapeutic group. Children have been seen to make outstanding progress in all aspects of their development.

## Conclusion

Child abuse is the result of a process with origins years, sometimes generations, before the event. The process is complex and different for every family. Factors in the parents' biographies, social problems, and ill health are all included. Identification of the syndrome needs recognition of the continuing process rather than diagnosis of an isolated medical event.

Most abusing families are known to the family doctor—firstly, because there is increased actual ill health and, secondly, because medical symptoms are often used as a way of seeking help. If the family doctor regards each consultation as part of the family dynamics and not as a single isolated event he has the unique opportunity for recognising early predictors of child abuse.

Early recognition of the problem is itself a step towards prevention. Reluctance to make the diagnosis could increase the risk. We have shown that the primary health care team can attempt to treat the problem of child abuse in the community and work towards prevention with the back-up of specialist services.

## References


1 Ounsted, C, Oppenheimer, R, and Lindsay, J, *Developmental Medicine and Child Neurology*, 1974, **16**, 447.
2 Lynch, M, Steinberg, D, and Ounsted, C, *British Medical Journal*, 1975, **2**, 127.
3 Lynch, M A, *Lancet*, 1975, **2**, 317.
4 Ounsted, C, and Lynch, M, *Child Abuse and Neglect. The Family and the Community*, ed R E Helfer and C H Kempe. Cambridge, Mass, Ballinger, 1976.
5 Lynch, M, and Ounsted, C, in *Child Abuse and Neglect. The Family and the Community*, ed R E Helfer and C H Kempe. Cambridge, Mass, Ballinger, 1976.
Lynch, M, Roberts, J, and Gordon, M, *Developmental Medicine and Child Neurology*. In press.
7 Department of Health and Social Security, *Memorandum on Non-accidental Injury to Children*. London, HMSO, 1974.
8 Department of Health and Social Security, *Non-accidental Injury to Children. Model Instructions for Accident and Emergency Departments*. London, HMSO, 1975.
9 Department of Health and Social Security, *Non-accidental Injury to Children: Reports from Area Review Committees*. London, HMSO, 1976.
10 Oliver, J E, and Taylor, A, *British Journal of Psychiatry*, 1971, **119**, 473.
11 Oliver, J E, and Cox, J, *British Journal of Psychiatry*, 1973, **123**, 81.
12 Ounsted, C, *World Medicine*, 1975, **10**, 27.
13 Browne, K, and Freeling, P, *The Doctor-Patient Relationship*, p 62. Edinburgh, Livingstone, 1967.
14 Gray, J, *et al*, *The Denver Predictive Study*. To be published.
15 Gray, J A, *Handbook of Psychopharmacology*, ed L Iverson, S Iverson, and S Snyder. New York, Plenum Press. In press.


# *Statistics at Square One*

# XX—Correlation (concluded)

T D V SWINSCOW

*British Medical Journal*, 1976, **2**, 802-803

### The regression equation

Correlation between two variables means that when one of them changes by a certain amount the other changes on the average by a certain amount. For instance, in Dr Green's children (Part XIX) greater height is associated on the average with greater anatomical dead space. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x. The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a *regression line* on a scatter diagram, and in the simplest case this is assumed

**British Medical Journal**
T D V SWINSCOW, MSC, MB, deputy editor

to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive), the slope is upwards from left to right; when one set decreases as the other increases, the slope is downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Apart from its being straight we have to define two other features of it if we are to place it correctly on the diagram. The first of these is its distance above the base line; the second is its slope. They are expressed in the following *regression equation:*

$$y = a + bx.$$

With this equation we can find a series of values of y, the dependent variable, that correspond to each of a series of values of x, the independent variable. The letters a and b are the *regression coefficients.* They have to be calculated from the data. The letter a signifies the distance above the base line at which the regression line cuts the vertical (y) axis. The letter b signifies the amount by which a change in x must be multiplied to give the corresponding average change in y. In this way it represents the degree to which the line slopes upwards or downwards.

Once the correlation coefficient has been computed the regression coefficients are easy to work out. We use results that we have already obtained. The formulae for finding a and b are as follows (in the order in which we calculate them):

$$b = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\Sigma(x-\bar{x})^2}. \qquad a = \bar{y} - b\,\bar{x}.$$

The calculation of the correlation coefficient in Part XIX on Dr Green's data gave the following:

$\Sigma(x-\bar{x})\,(y-\bar{y}) = 5\,426{\cdot}6$
$\Sigma(x-\bar{x})^2 = 5\,251{\cdot}6$
$\bar{y} = 66{\cdot}93$
$\bar{x} = 144{\cdot}6.$

Applying these figures to the formulae for the regression coefficients, we have:

$$b = \frac{5\,426{\cdot}6}{5\,251{\cdot}6} = 1{\cdot}033$$
$$a = 66{\cdot}93 - (1{\cdot}033 \times 144{\cdot}6) = -82{\cdot}4.$$

Therefore the equation for the regression of y on x becomes in this case

$$y = -82{\cdot}4 + 1{\cdot}033x.$$

This means that, on the average, for every increase in height of 1 cm the increase in anatomical dead space is 1·033 ml *over the range of measurements made.*

The line representing the equation is shown superimposed on the scatter diagram of Dr Green's data in fig 20.1. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle, and one on the right, and substitute these in the equation. Dr Green's figures come out as follows:

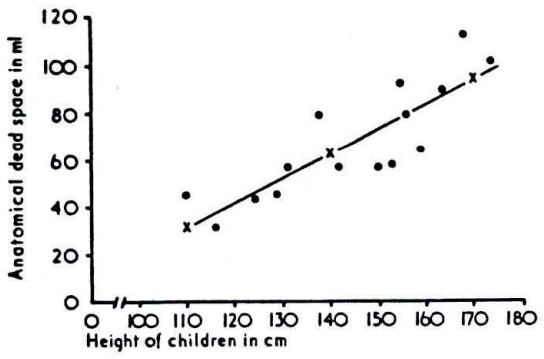



FIG 20.1—Regression line drawn on scatter diagram relating height and pulmonary anatomical dead space in 15 children (fig 18.2).

If x = 110, $y = (1{\cdot}033 \times 110) - 82{\cdot}4 = 31{\cdot}2$
If x = 140, $y = (1{\cdot}033 \times 140) - 82{\cdot}4 = 62{\cdot}2$
If x = 170, $y = (1{\cdot}033 \times 170) - 82{\cdot}4 = 93{\cdot}2$

Though two points are enough to define the line, three are better as a check. Having put them on the scatter diagram, we simply draw the line through them.

Regression lines give us useful information about the data they are collected from. They show how one variable changes on the average with another, and they can be used to find out what one variable is likely to be when we know the other—provided we ask this question within the limits of the scatter diagram. But to project the line at either end—to extrapolate—is always risky. The relationship between x and y may change, or some kind of cut-off point may exist. For instance, a regression line might be drawn relating the chronological age of some children to their bone age, and it might be a straight line from, say, the age of 5 years to 10, but to project it up to the age of 30 would clearly lead to error.

*Exercise 20.* From the data in exercise 19: if values of x represent mean distances of nearest five termite hills and values of y represent percentages of kala-azar cases, what is the equation for the regression of y on x? What does it mean? *Answer:* $y = 36 - 0{\cdot}23x$. It means that, on the average, for every 10 m increase in mean distance of the five nearest termite hills, the percentage of cases of kala-azar falls by 2·3 ($= 10 \times 0{\cdot}23$). This can be safely accepted only within the area measured here.

---

This article concludes the series "Statistics at Square One." The articles are being revised for republication shortly in book form. Two further sections will then be added, on rank correlation and Fisher's exact probability test.

---

*Would fructose at the end of a day reduce the length of time in which ethanol would be exercising its toxic effect on tissue, particularly nervous tissue? How much fructose would be required to metabolise a given quantity of alcohol?*

Undoubtedly fructose taken at the end of the day would accelerate the metabolism of alcohol circulating then. This does not necessarily mean that the hangover effects of alcoholic drinks could be avoided, because these are partly caused by substances other than ethyl alcohol. At least 25 to 50 g of fructose would be needed and this would make a small but perhaps appreciable contribution to daily calorie intake.

*Is it possible to diagnose conditions of insidious onset such as neoplasms of the pancreas, kidney, and central nervous system at an early and treatable stage by routine ultrasound scanning and computerised tomography?*

There is no simple answer to this complex question. The short answer is No, not until "practical possibility," "early and treatable stage," and "routine," have been defined accurately. The reasons for this are: (*a*) ultrasound has no place in detecting central nervous system tumours compared with the pre-eminence of computerised tomography; (*b*) excretion urography is still the first choice for investigating the urinary tract, and subsequent investigations depend entirely on the results of the urogram and not on a predetermined routine. Both ultrasound and computerised tomography are important complementary tests, especially in distinguishing cysts from solid tumours; (*c*) the pancreas and other upper abdominal retroperitoneal structures are highly promising for computerised tomography, and advances have been made in grey-scale ultrasonic scanning of this region as well. The resolution of both techniques is good, but pancreatic pseudocyst is recognised more readily than carcinoma by both techniques and these striking developments still require correlation with other techniques before their place can be judged; and (*d*) the concept of having "routine" ultrasound scanning and computerised tomography is not acceptable until the method of patient selection has been defined. The potential demands for computerised tomography exceed the capacity of the instruments available, and there is no question of conducting population surveys.

*Journal of Tropical Medicine and Hygiene* 1983, **86**, 37–46

# High prevalence of eye disease in a Haitian locale

David A. Newsome[1], Roy C. Milton[2] and Gerard Frederique[3]

[1]*Section on Retinal and Ocular Connective Tissue Diseases, Clinical Branch, and the* [2]*Office of Biometry and Epidemiology, National Eye Institute, National Institutes of Health, Bethesda, Maryland, and the* [3]*Departments of Opthalmology, University of Haiti, Port-au-Prince, and Georgetown University Medical Center, Washington, DC, USA*

## Introduction

Haiti occupies the western one-third of the Caribbean island of Hispanola. With one of the very lowest socioeconomic levels in the western hemisphere, it is not surprising that the prevalence of disease is considerable, and that access to appropriate care is limited. Such obvious need among the Haitian people has stimulated assistance efforts on the part of various international organizations. Ocular diseases are particularly pressing: problems related to hypovitaminosis A and protein energy malnutrition have received considerable attention (WHO 1973, Sommer *et al.* 1976). The high prevalence of other diseases has been supported in the past only by anecdotal information.

A locality in the southern peninsula of Haiti has been designated for the establishment of the first fulltime ophthalmic service in that area. It is important to know the scope and severity of problems in an area to be served, to plan effective intervention and evaluate the results. No reliable data exist on general ophthalmic disease prevalence in Haiti. Such information should provide guides for an ophthalmic service, as well as aid in the further development of a community-based maternal and child health program.

We report here the results of a survey of the predicted service area of a planned ophthalmic clinic and dispensary in Leogane, a village about 45 km west of the Haitian capital, Port-au-Prince. Our data confirm a high prevalence of eye disease in general and of certain diseases in particular. These data document for the first time the scope of general ophthalmic problems in a Haitian population.

Correspondence to: Dr David A. Newsome, The Wilmer Ophthalmological Institute, Johns Hopkins Hospital, Baltimore, Maryland 21205, USA.

## Materials and methods

### SURVEY AREA

The facility to be constructed will be in the small town of Leogane. We selected a survey area within 10 km of Leogane and determined by natural boundaries, such as mountains, and with varied accessibility. The only previous ophthalmic service in the survey area had been provided by irregular occasional visits of United States ophthalmologists to Hospital Ste. Croix, Leogane. Reliable demographic and census data for the survey area were sparse. The enumeration of dwellings developed for the national malaria eradication programme yielded an estimate of about 5000 dwellings in the survey area and a population of about 50 000.

The survey area included coastal areas, plains and mountains, with 90% rural and 10% urban population. Good all-weather roads cross the survey area coming from Port-au-Prince and running south from Leogane. Much of the region is accessible only by unimproved roads or by foot.

The survey area included sugar cane plantations as well as mountainous areas. Many families in the coastal and plains regions participate in the cash household economy of sugar cane workers. In the mountainous area families depend on household gardens and animal husbandry in addition to some cash economic elements.

### SURVEY DESIGN

A 5% sample of the survey area was chosen based on both statistical and logistical consider-

ations. Survey sites were selected in proportion to the area of each geographic region (seaside, plains, mountainous, highway), maintaining where possible the representation of various types of population concentration (small town, closely grouped dwellings, scattered dwellings), accessibility to four-wheel drive vehicles (good, fair, poor) and recent or current health-care activity (hospital in vicinity, health centre, maternal and child health programme, no neighbourhood care). Each of the 15 survey sites comprised two separate clusters of groups of contiguous households, with an average cluster having about 85 persons of all ages.

### SURVEY TEAMS AND PLAN

The enumeration team visited the survey area during September, 1979. They confirmed and amended dwelling maps and made preliminary recommendations concerning sample sites. The Haitian project co-ordinator, in consultation with the United States project supervisor, made final cluster selections and the field schedule. The actual survey was conducted during October 1979. A sample access team announced and explained the purpose of the survey as a general medical examination to households in the selected cluster the day before the examination. Enumeration data were confirmed.

The field team had three major components. The sample access team, using enumeration data, brought cluster householders to the examination site. The reception and data collection team interviewed and created a data card for each survey participant. The examination team consisted of one Haitian ophthalmologist, one Haitian paediatrician, two United States optometrists with graduate training in public-health techniques, and four Haitian assistants. Local officials and leaders assisted in pre-examination contact as well as bringing cluster households to the survey team. Local co-operation was excellent and we estimate that over 90% of the potential survey participants were available for examination even though some wage earners were absent.

### EXAMINATION PROTOCOL

All team members had standardization training and were familiar with the protocol which defined specific criteria for recording individual findings. Visual acuity (with glasses if available) was tested at 6 m by using an illiterate E chart in indirect sunlight. The results were recorded according to WHO guidelines (WHO 1980). Examinations were conducted in an enclosed area (e.g. church, school) in reduced illumination. Cataract was said to be present when, by observation with loupes and handlight through the undilated pupil, central opacities (whitish in colour) or brunescent changes were observed. The presence of these changes was confirmed by direct ophthalmoscopic observations. It should be noted that this technique will detect definite changes in the central lens, but will not readily identify peripheral cortical changes or mild nuclear sclerosis.

Intraocular pressure was determined with a standard hand-held indentation tonometer. Fundoscopy was performed by direct ophthalmoscopy. Subjects with positive findings were re-examined by at least one other senior team member. In cases of corneal scarring additional specific history (e.g. nutritional status, infections, trauma) was sought and recorded. General physical conditions of significance (e.g. kwashiorkor) were noted. Subjects with ocular problems requiring follow-up were either treated on the spot or referred to clinics held specifically for such persons. (The Dominique-Coicou Eye Care Clinic, staffed fulltime by the Georgetown University Department of Ophthalmology senior resident and supervised by a United States board-certified Haitian ophthalmologist, was opened in February 1980, and now provides regular service to the survey area.) The team was equipped with emergency first aid and surgical supplies, milk rations and vitamin A capsules in accordance with WHO recommendations.

### DATA RECORDING, PROCESSING AND ANALYSIS

The data for each person were recorded on an individual card designed for simple coding, keypunching and computer processing. Questions were asked to determine each subject's perceived need for eye care in the preceding year, whether help had been obtained and from what source (traditional, doctor, Haitian physician, United States volunteer physician) and whether

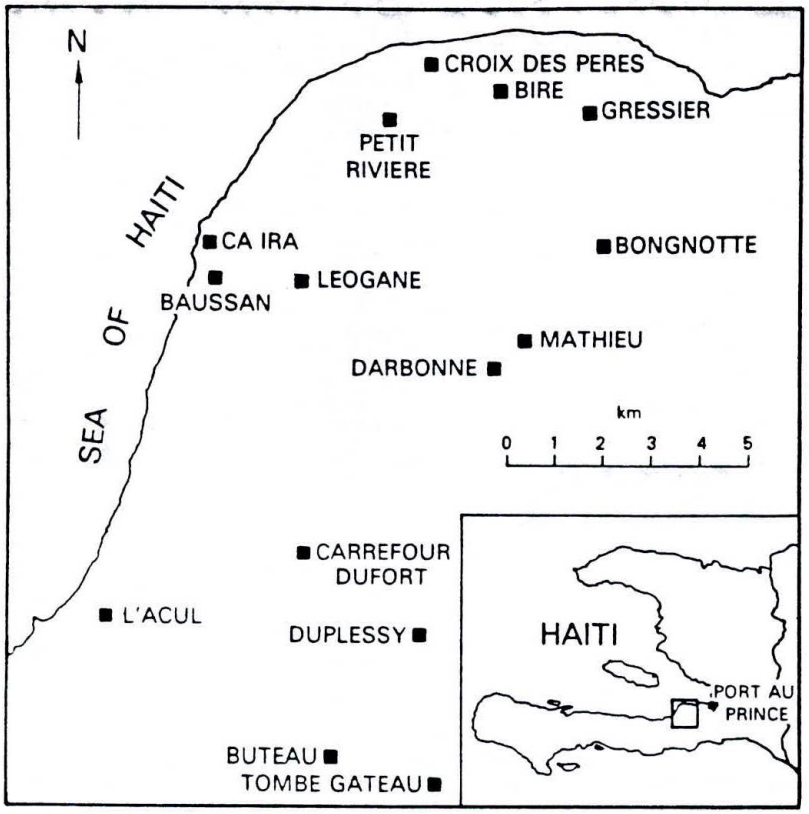



**Figure 1.** Location of the 15 sample sites in the Leogane area.

the subject wore eyeglasses (adult householders supplied the information for children). Anatomic and examination findings were similarly recorded. Prevalence of ocluar conditions is reported by age–sex groups.

## Results

### POPULATION CHARACTERISTICS

The sample sites were scattered throughout the 90% rural target area (Figure 1). Characteristics of the sites are given in Table I. The total number of persons examined at each site varied from 137 at Ca Ira to 210 at Bire. A total of 2531 persons were examined, about 4.5% of the estimated base population. The distribution of the population examined according to age and sex is shown in Table 2 and is compared in Figure 2 with the 1975 national distribution found by the Haitian Department de Santé Publique. The main difference between these two age pyramids is the somewhat lower percentage of children in the present survey. This and other differences observed may be due to regional or sampling variation. As with other age–sex distributions of various Haitian populations surveyed on other occasions, our distribution shows a slight predominance of women.

### SUBJECTIVE PERCEPTION OF EYE PROBLEMS AND PATTERNS OF THE USE OF MEDICAL CARE FOR THOSE PROBLEMS

Of the 2531 persons surveyed, 1900 (75%) responded affirmatively to the question of whether they had serious eye problems within the past year. In contrast, only 468 persons (18%) had seen a doctor for their eyes during this period. Almost 60% of those doctor–patient encounters for ocular problems included a Haitian physician. The other encounters presumably involved one of a number of voluntary non-Haitian physicians. One indicator of the

**Table 1.** Characteristics of survey sample sites

| Sample site | Population[a] | No. Examined | Accessibility by four-wheel drive vehicle | Type of health care available |
|---|---|---|---|---|
| Town | | | | |
| Leogane | 5416 | 199 | Good | Hospital |
| Seaside | | | | |
| Ca Ira | 1526 | 137 | Good | None |
| Baussan | 6325 | 147 | Fair | None |
| Plains | | | | |
| Croix des Peres | *b* | 161 | Fair | None |
| Gressier | 3375 | 205 | Fair | Maternal-child health programme |
| Bire | 1419 | 210 | Fair | None |
| Petit Riviere | 8847 | 170 | Fair | None |
| Bongnotte | 2175 | 157 | Poor | None |
| Darbonne | 3444 | 150 | Good | Visiting eye doctor |
| Mathieu | 4656 | 168 | Fair | Maternal-child health programme |
| Highway village | | | | |
| Carrefour Dufort | 3720 | 178 | Good | Dispensary |
| L'Acul | 6569 | 194 | Good | Maternal-child health programme |
| Mountain: | | | | |
| Buteau | 2075 | 175 | Good | Maternal-child health programme |
| Duplessy | 3388 | 139 | Poor | Dispensary |
| Tombe Gateau | 3632 | 141 | Good | Dispensary |
| Total | 56 567 | 2531 | | |

[a]Estimated from the National Survey for Malaria Eradication.
[b]The population of Petite Riviere includes Croix des Peres.

**Table 2.** Distribution of persons examined, by age and sex

| Age (years) | Male | Female | Total |
|---|---|---|---|
| 0–4 | 96 | 116 | 212 |
| 5–9 | 82 | 103 | 185 |
| 10–14 | 109 | 100 | 209 |
| 15–19 | 89 | 153 | 242 |
| 20–24 | 100 | 161 | 261 |
| 25–29 | 55 | 124 | 179 |
| 30–34 | 52 | 93 | 145 |
| 35–39 | 60 | 78 | 138 |
| 40–44 | 74 | 121 | 195 |
| 46–49 | 72 | 84 | 156 |
| 50–54 | 90 | 99 | 189 |
| 55–59 | 47 | 51 | 98 |
| 60–64 | 40 | 79 | 119 |
| 65–69 | 35 | 39 | 74 |
| 70+ | 59 | 70 | 129 |
| Total | 1060 | 1471 | 2531 |

socioeconomic and ophthalmic service level of the survey population is that only 7% (182) of the persons surveyed had eyeglasses.

VISUAL ACUITIES

Examination for visual acuity was conducted with the E chart at 6 m. If the visual acuity was not sufficient to be measured on the chart, it was then estimated in terms of finger counting, hand movements and amount of light perception. It was usually impossible to test the vision accurately in the 0–4 year old group. However, if significant pathology was noted, the examiner attempted to estimate its effect on visual acuity. Prevalence of low visual acuity by age and sex is given in Table 3. The prevalence of visual acuity $\leqslant 6/60$ rose sharply after age 50 years to over 31% in the 70+ year old age group. The prevalence of this level of visual acuity did not differ significantly by sex. Prevalence of visual acuity in the better eye $\leqslant 6/30$ and $\leqslant 6/60$ was 7.3 and 3.7% respectively. The major causes of significantly decreased visual acuity included cataract, pterygium, corneal scarring and glaucoma.

Three categories of problems causing ex-

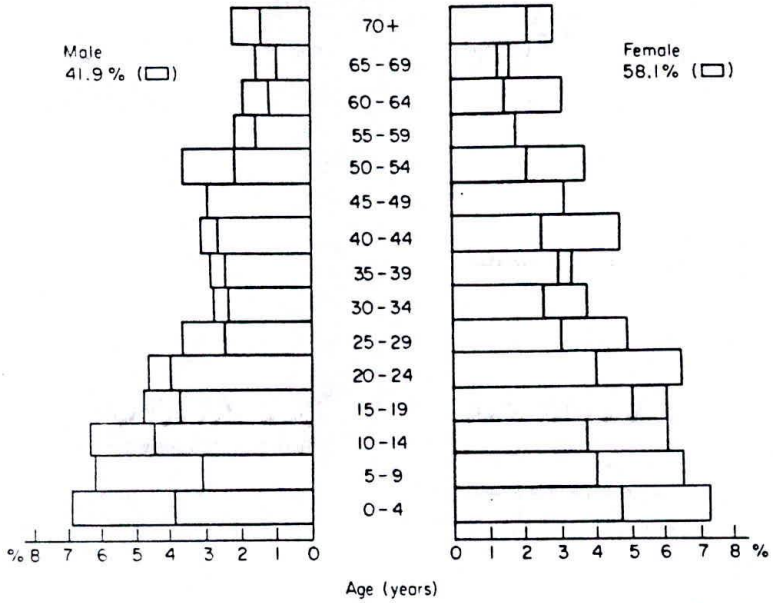



**Figure 2.** Percentage distribution of the sample population (present survey 1979, ▨) and reference population (national survey 1975, □), by age and sex.

**Table 3.** Percentage[a] of persons with low visual acuity, by age and sex

| | Visual acuity in the better eye[b] | | | | | |
|---|---|---|---|---|---|---|
| | ⩽ 6/60 | | | ⩽ 6/30 | | |
| Age (years) | Male | Female | Total | Male | Female | Total |
| 0–9 | 0.6 | 0.0 | 0.3 | 0.6 | 0.0 | 0.3 |
| 10–19 | 1.0 | 0.4 | 0.7 | 2.0 | 0.8 | 1.3 |
| 20–29 | 1.3 | 1.1 | 1.1 | 1.9 | 1.4 | 1.6 |
| 30–39 | 0.0 | 1.8 | 1.1 | 3.6 | 2.9 | 3.2 |
| 40–49 | 2.7 | 2.4 | 2.6 | 4.1 | 7.8 | 6.3 |
| 50–59 | 2.2 | 5.3 | 3.8 | 6.6 | 16.7 | 11.8 |
| 60–69 | 10.7 | 11.0 | 10.9 | 21.3 | 27.1 | 24.9 |
| 70+ | 35.6 | 28.6 | 31.8 | 45.8 | 42.9 | 44.2 |
| Total | 3.9 | 3.6 | 3.7 | 6.6 | 7.7 | 7.3 |

[a]Percentage of age–group.
[b]With glasses if available.

tremely reduced vision or blindness are of special interest. The survey included one 3 year old child with retinoblastoma, seven person with enucleation of one or both eyes (5 persons with a history of trauma, 2 with a history of keratomalacia) and 12 persons with retinal or optic nerve disease.

CATARACT

Cataracts of all types were observed in 6.2% of the total population (Table 4). Congenital cataract as diagnosed by history and anatomical appearance was found in 0.8% of the 0–9 years age group. Cataracts of all types were increasingly common over the age of 40 years, with

**Table 4.** Percentage[a] of persons with cataract of all types, by age and sex

| Age (years) | Male | Female | Total |
|---|---|---|---|
| 0–29 | 0.6 | 0.3 | 0.4 |
| 30–39 | 0.8 | 1.2 | 1.1 |
| 40–49 | 1.4 | 3.4 | 2.6 |
| 50–59 | 8.8 | 10.9 | 9.8 |
| 60–69 | 20.0 | 30.5 | 26.4 |
| 70+ | 49.2 | 44.3 | 46.5 |
| Total | 5.8 | 6.4 | 6.2 |

[a]Percentage of age–sex group.

prevalence 2.6% for ages 40–49, 9.8% for ages 50–59, 26.4% for ages 60–69 and 46.5% for ages 70 years and over. Men and women did not differ significantly in prevalence of cataract. Aphakia was rare (5 persons). Cataract uncomplicated by pterygium and with visual acuity $\leqslant 6/30$ was found in 1.3% of the total population and 8.5% in the 70+ years age group.

PTERYGIUM AND PINGUECULUM

Pterygium was not seen below 10 years of age, was rare until age 20 and the increased with age to almost 70% for ages 70 years and over (Figure 3). Pterygium involving the pupillary zone and accompanied by reduced visual acuity ($\leqslant 6/30$) was not observed in men but was seen in 0.8% of women aged 30 years and over. Pingueculum, which has been thought to be the lesion which can give rise to pterygium, was rare through age 14 years. The prevalence of pingueculum increased to 26% for ages 30–49 years and then decreased (Figure 3).

CONJUNCTIVITIS

Conjunctivitis of all types was present in all age groups (Figure 3), but was more common (10.8% prevalence) among children. Bacterial conjuctivitis was the form most frequently encountered, with a significant number of cases of vernal and presumed viral conjunctivitis also seen. The early stages of trachoma were considered in the differential diagnosis but no definite cases were detected. Appropriate topical therapy was dispensed for these cases.

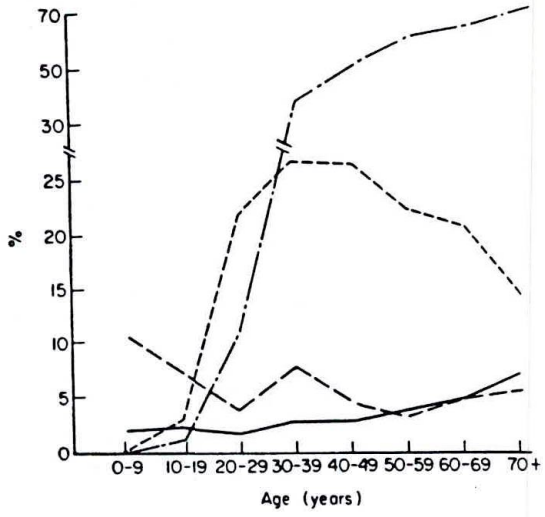



**Figure 3.** Percentage distribution, by age, of conjunctivitis (——), corneal scars (——), pingueculum (– – – –) and pterygium (— — —).

GLAUCOMA

For the purpose of our survey, glaucoma was said to be present when an intraocular pressure $\geqslant 25$ mmHg was observed in at least one eye and verified by a second examiner. Glaucomatous disc changes were commonly observed with these elevated pressures. Prevalence of glaucoma by this definition for all ages was 5.8% (Table 5), with no apparent differences by sex. Noticeable elevation of intraocular pressure began in the 30–39 year age group (2.8%), steadily increasing with every decade thereafter to nearly 28% of persons 70 years of age and over. Significantly elevated intraocular pressure was the sole abnormality detected in a substantial number of persons with low visual acuity.

**Table 5.** Percentage[a] of persons with elevated intraocular pressure (IOP $\geqslant 25$ mmHg), by age and sex

| Age (years) | Male | Female | Total |
|---|---|---|---|
| 0–29 | 0.2 | 0.3 | 0.2 |
| 30–39 | 4.5 | 1.8 | 2.8 |
| 40–49 | 10.3 | 10.2 | 10.3 |
| 50–59 | 13.7 | 11.3 | 12.5 |
| 60–69 | 17.3 | 11.9 | 14.0 |
| 70+ | 22.0 | 32.9 | 27.9 |
| Total | 6.2 | 5.4 | 5.8 |

[a]Percentage of age–sex group.
IOP, Intraocular pressure.

**Table 6.** Causes of corneal scarring

| Cause | No. of cases | No. with visual axis involved |
|---|---|---|
| Trauma | 10 | 8 |
| Vernal conjunctivitis | 10 | 2 |
| Nutrition | 5 | 5 |
| Infections (bacterial, viral) | 2 | 2 |
| Miscellaneous known | 17 | 10 |
| Uncertain | 16 | 10 |
| Total | 60 | 37 |

Absolute glaucoma alone was the most frequent explanation for low-light perception eyes. Only one case of congenital glaucoma was observed.

CORNEAL SCARS OF ALL TYPES

Corneal scarring was distributed throughout all age groups, with prevalence increasing slightly with age (Figure 3). Corneal scarring can be caused by a variety of pathologic processes. It was not possible in all of the cases of corneal scarring identified in this survey to assign an etiology with certainty. However, in those cases where an underlying cause could be determined with some certainty by history or associated findings, trauma, vernal conjunctivitis and nutritional causes were frequently encountered (Table 6). The 'uncertain' category contains six persons whose corneal scars clinically were most probably due to previous nutritional eye disease. The 'miscellaneous' group includes such problems as severe trichiasis, keratoconus and pterygium. The combination of trichiasis and corneal scarring as seen in eight persons was likely due to previously active trachoma.

GEOGRAPHIC VARIATION OF SPECIFIC EYE FINDINGS

The standardized prevalence of selected ocular disease findings and poor visual acuity by survey site is given in Table 7, showing noticeable differences depending on the geographical location of the survey site. Prevalence of separate or coexisting eye conditions was generally lower in individual sites in the mountainous and highway

**Table 7.** Age–sex standardized[a] percentage of persons with selected characteristics, by sample site

| Sample site | (1) VA 6/60 or worse | (2) Glaucoma: IOP ⩾25 mmHG[b] | (3) Corneal scarring[b] | (4) Cataract[b] | (5) At least one of (1)–(4) |
|---|---|---|---|---|---|
| Total, all sites | 3.7 | 5.8 | 2.8 | 6.2 | 13.0 |
| Town | | | | | |
| Leogane | 3.1 | 8.0 | 3.7 | 6.3 | 16.3 |
| Seaside (Total) | (5.4) | (5.7) | (4.7) | (9.6) | (17.3) |
| Ca Ira | 4.4 | 7.6 | 6.0 | 6.9 | 19.0 |
| Baussan | 6.6 | 2.7 | 2.3 | 12.2 | 14.5 |
| Plains (Total) | (4.2) | (6.2) | (3.9) | (7.7) | (15.6) |
| Croix des Peres | 6.8 | 9.1 | 6.8 | 9.0 | 23.6 |
| Gressier | 5.3 | 6.8 | 3.5 | 8.0 | 13.8 |
| Bire | 3.6 | 4.7 | 2.7 | 8.0 | 13.3 |
| Petit Riviere | 0.5 | 4.0 | 2.4 | 6.6 | 11.3 |
| Bongnotte | 3.1 | 5.9 | 3.6 | 8.0 | 11.8 |
| Darbonne | 6.7 | 7.7 | 3.5 | 5.9 | 15.9 |
| Mathieu | 3.6 | 6.7 | 7.0 | 7.5 | 20.5 |
| Highway village (Total) | (5.0) | (4.0) | (0.2) | (5.5) | (8.1) |
| Carrefour Dufort | 6.0 | 3.9 | 0.3 | 6.7 | 9.1 |
| L'Acul | 2.7 | 2.7 | 0.0 | 2.7 | 4.9 |
| Mountain (Total) | (0.6) | (3.9) | (0.8) | (1.1) | (5.6) |
| Buteau | 1.0 | 2.7 | 1.4 | 2.1 | 6.2 |
| Duplessy | 2.3 | 4.4 | 1.2 | 2.3 | 8.5 |
| Tombe Gateau | 0.0 | 3.1 | 0.0 | 0.0 | 3.1 |

[a]Standard population is total number examined, by sex and 10-year age groups.
[b]In at least one eye.
VA, Visual acuity; IOP, intraocular pressure.

Table 8. Percentage[a] of persons with blindness, by age, definition of blindness and eye disease occurring with the reduced vision

| Age (years) | No. of persons | Eye disease occurring with reduced vision: Total | Cataract | Glaucoma | Corneal scarring[b] | Pterygium[b] | Other |
|---|---|---|---|---|---|---|---|
| Visual acuity in the better eye[c] ⩽ 6/60 | | | | | | | |
| < 40 | 1571 | 0.76 | 0.13 | 0.00 | 0.19 | 0.06 | 0.38 |
| 40+ | 960 | 8.54 | 5.52 | 3.75 | 0.83 | 0.31 | 1.56 |
| Total | 2531 | 3.71 | 2.17 | 1.42 | 0.43 | 0.16 | 0.83 |
| Visual acuity in the better eye[c] ⩽ hand motion | | | | | | | |
| < 40 | 1571 | 0.06 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 |
| 40 + | 960 | 1.88 | 1.15 | 1.25 | 0.31 | 0.10 | 0.21 |
| Total | 2531 | 0.75 | 0.43 | 0.47 | 0.16 | 0.04 | 0.08 |

[a]Percentage of age group.
[b]Involving visual axis.
[c]With glasses if available.

regions. For these regions, standardized prevalence of serious eye problems (category 5: at least 1 of poor vision, glaucoma, corneal scarring and cataract) was one-third to one-half of that in the other regions. Reduction in prevalence of poor vision, corneal scarring, and of cataract in the mountainous region was more striking, to generally 20% or less of the values for lower-lying areas.

BLINDNESS

Prevalence of low visual acuity (⩽6/60 in the better eye) as shown in Table 3 would be appropriate for comparisons with legal blindness as commonly defined in developed countries. Different priorities and allocation of resources in developing countries may suggest a more stringent criterion for blindness, such as visual acuity of hand motion or worse in the better eye. The observed frequency of blindness (acuity hand motion or worse in the better eye) and low visual acuity (⩽6/60 in the better eye) occuring with cataract, glaucoma, corneal scarring involving the visual axis, and pterygium involving the pupillary zone is given in Table 8.

The observed blindness rate for hand motion or worse is 19 per 1000 persons aged 40 years and over and 7.5 per 1000 among all persons. About one-half of these blind persons had both cataract and glaucoma. Blindness defined by vision ⩽6/60 was 85 per 1000 persons aged 40 years and over and 37 per 1000 among all persons. By this definition cataract and glaucoma occurred together in 30% of blind persons and about 30% were associated with neither cataract nor glaucoma.

## Discussion

Cataracts, corneal scarring, glaucoma and pterygium were associated with severely reduced vision or blindness in our survey area, occurring at prevalence levels generally greater than those found in the United States or other countries in the western hemisphere. While the prevalence of these conditions increased with age, glaucoma appeared at younger ages than is usually reported in other populations. The majority of serious eye problems observed in persons with low visual acuity are usually amenable to therapeutic intervention by both surgical and non-surgical means.

Chronic glaucoma in the United States, especially among whites, is usually associated with an older age group. Chronic glaucoma with pressure elevation and frequently severe disc changes is a disease of younger adults in the Haitian population studied. This finding parallels observations made in populations in Africa and from work with populations of African origin (Newmann & Zauberman 1965, Hiller & Kahn 1975, Amoni 1980, Coulehan *et al.* 1980).

The large contribution of African origin to the genetic make-up of the Haitian population may in part explain the high prevalence of glaucoma at relatively early ages in Haiti. The prevalence of intraocular pressure $\geqslant$25 mmHg for persons aged 50 years and over was 16%, markedly greater than the 3.6% prevalence in a recent survey of a white population in the United States (Kahn *et al.* 1977). We also encountered a number of patients in the under-20 years age group who had elevated intraocular pressure readings along with disc changes. It is possible that this much smaller group represents a different genetic variant of the adult disease. It is widely appreciated that chronic glaucoma causes few symptoms until advanced stages of the disease, emphasizing the need for screening and primary eye-care delivery to populations with such high prevalences as we encountered in the Leogane area.

Cataract as observed by the methods of this survey was usually nuclear sclerosis, frequently accompanied by pterygium. Cataract prevalence for persons aged 50 years and over (from Tables 2, 4) was 23% comparable with the prevalence of nuclear sclerosis or aphakia (26.5%) in those of similar age in a United States study (Leibowitz *et al.* 1980). Comparision of the two surveys with respect to prevalence when concurrent significantly reduced vision is added as a criterion suggests a twofold excess in the Haitian prevalence, but this excess may be due in part to the frequent occurrence of pterygium with cataract.

It is likely that a number of factors other than heredity influence the prevalence of serious eye problems in our survey area. The sites surveyed were located in identifiable geographic areas, with each area rather homogeneous with respect to household economics and dietary habits. For example, communities along the railroad tracks near the sugar cane fields have a largely cash household economy while those in more mountainous areas more often have household gardens and raise animals for consumption.

The level of overall prevalence of serious eye problems by sample site was remarkably consistent within each geographic area. The town, seaside and plains areas had the highest prevalence of affected individuals. The reduction in serious eye findings encountered in the mountainous area as compared with the lower survey sites, while possibly related to existing health-care facilities, may also be related to different diet and environmental factors, though genetic factors may play a part. There was no evidence of selective migration from the mountainous area that might reduce prevalance there and the age distributions of the populations of the five geographic areas were quite similar.

Nutritional status in our survey can only be roughly inferred from age. The number of children encountered with corneal scarring (5) likely due to hypovitaminosis A, protein energy malnutrition and active xerophthalmia (1) indicated that vitamin A–protein energy malnutrition may be classified as a public health problem in the survey area according to World Health Organization criteria. Hypovitaminosis A is the 'major cause' of bilateral blindness among Haitian children (Sommer *et al.* 1976). Our investigations showed 1% prevalence (5) of corneal scarring of presumed nutritional origin among children in southern Haiti $<$10 years old. This prevalence is the same as that found by Sommer *et al.* (1976) in a country-wide prevalence survey in Haiti in 1976, but much higher than that reported in southern Haiti (0.12%). We saw no Bitot's spots, confirming the observation by Sommer *et al.* (1976) that Bitot's spots are at least uncommon if not extremely rare in the Haitian population in spite of nutritional status. This finding casts further doubt on the reliability of the detection of Bitot's of Bitot's spots as an indicator of vitamin A nutritional status. Our experience with the 60 persons encountered in the survey with corneal scarring confirms the uncertainties described by others in determining retrospectively the aetiologies of corneal scarring. Despite out best attempts to elicit a reliable history of malnutrition, measles, other infections and other associated causes, including trauma, we could not establish a definite aetiology in nearly 25% of the cases and no reliable medical records were available. According to WHO and to the International Vitamin A Consultative Group, corneal scarring is felt to be an important indicator of severe nutritional eye disease in a particular area. Before assigning even one case to be a category which may in-

fluence strongly the interpretation of study results, data from multiple observers and rigorous criteria for the retrospective assignment of an aetiological diagnosis were used.

Trachoma is known from clinical experience to be endemic in Haiti and we encountered a substantial number of persons with disabling sequelae of, presumably, that disease. We did not definitely detect active cases, although in the field early disease can be incorrectly categorized. With that caution, it would appear that trachoma is endemic at a low level in the study area. Records from 2 years (1980, 1981) of patient visits to a clinic opened in the survey area confirm the low prevalence of active trachoma (G. Frederique, unpublished work).

Our study has established the scope and magnitude of general ophthalmic disease in a particular Haitian locality. Our findings confirm the high prevalence of severe ocular disorders in this population and highlight the need for primary eye-care delivery. The coming of a fulltime ophthalmic service to this area should have substantial impact to improve the general level of eye health in this community. Most of the severe eye problems including glaucoma, cataracts, visually disabling pterygium and diseases associated with corneal scarring could be helped by appropriate ophthalmic care. It seems reasonable to conclude that programs increasing the accessibility of ophthalmic care and increasing the public awareness of preventive eye care and of good nutrition may help to reduce the burden of blindness.

## References

Amoni D. D. (1980) Pattern of prevention of glaucoma in Kaduna, Nigera. *Glaucoma* **2**, 445

Coulehan J. L., Helzlsouer K. J., Rogers K. D. & Brown S. I. (1980) Racial differences in intraocular tension and glaucoma surgery. *American Journal of Epidemiology* **111**, 759

Hiller R. & Kahn H. A. (1975) Blindness from glaucoma. *American Journal of Ophthalmology* **80**, 62

Kahn H. A., Leibowitz H. M., Ganley J. P., *et al.* (1977) The Framingham Eye Study. I. Outline and major prevalence findings. *American Journal of Epidemiology* **106**, 17

Leibowitz H. M., Krueger D. E., Maunder L. R., *et al.* (1980) The Framingham Eye Study Monograph. *Survey of Ophthalmology* **24** (Suppl), 335

Neumann E. & Zauberman H. (1965) Glaucoma survey in Liberia. *American Journal of Ophthalmology* **59**, 8

Sommer A., Toureau S., Cornet P., *et al.* (1976) Xerophthalmia and anterior segment blindness. *American Journal of Ophthalmology* **82**, 439

World Health Organization (1973) *The Prevention of Blindness* Technical Report Series No. 518, WHO, Geneva

World Health Organization (1980) *Methods of Assessment of Avoidable Blindness* Offset Publication No. 54, WHO, Geneva