WHO/CDS/CPE/SMT/2000.2 Rev.1
Part I

Introduction to basic epidemiology and principles of statistics for tropical diseases control

Learner's Guide

Communicable Diseases Cluster
Department of Control, Prevention and Eradication
Social Mobilization and Training Unit
Updated July 2002

Trial Edition

The designations employed and the presentation of the material in this health information product do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by the World Health Organization in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

The World Health Organization does not warrant that the information contained in this health information product is complete and correct and shall not be liable for any damages incurred as a result of its use.

Table of contents

Learning Units

Foreword

This module uses a problem-solving approach to facilitate the learning of some basic epidemiological concepts and practices and simple statistics. It is designed for health workers responsible for tropical diseases control. It is considered to be fundamental to the learning of the epidemiological approach to malaria control and for a situation analysis.

The module is designed to stimulate active learning. The structure of the module can be seen from the table of contents. It is divided into two parts: *Learner's Guide* (Part I) and *Tutor's Guide* (Part II). The *Learner's Guide* contains basic information and exercises. The *Tutor's Guide* contains suggestions for using this training module and provides suggested answers to the exercises.

The module can be used in different ways. It is primarily created for group work as one element of a comprehensive training course on basic malariology and planning malaria control. The module can also be used separately for epidemiology or in-service training at appropriate levels of the health services. It can be a useful component of any programme for training in tropical disease control at the district and national levels. The *Learner's Guide* can be used for individual work. The *Tutor's Guide* is intended to supplement the tutor's own knowledge and experience and to guide the facilitators, or, in the case of individual study, to be used as an answer book. The training using this module is designed to be accomplished in 26 hours including one hour each for the pre and post tests (See Part II Introduction for a proposed timetable).

The module has been developed by Dr N. Binkin whilst on assignment to the International Course for Primary Health Care Managers and the Laboratory of Epidemiology and Biostatistics at the *Istituto Superiore di Sanità* in Rome. Dr Binkin was at the time a WHO staff member of the Division of Strengthening of Health Services, HQ, Geneva. Thanks go to Dr P.F. Beales, the late Dr R.L Kouznetsov and Dr F. A. Rio for reviewing the document and reorganizing it into a two-volume training module. The technical editing of the module was undertaken by Dr F.A. Rio and pursued by Dr M. Aregawi and Dr M.C. Thuriaux, all at WHO Headquarters. Participants and facilitators in English-, French- and Portuguese-language courses in Nazareth (Ethiopia), Cotonou and Maputo respectively have contributed greatly to the practical development of the module, which was supported by a financial contribution from the World Bank.

Introduction

A basic understanding of the principles of epidemiology and a knowledge of some simple statistics, and how and when to apply them are essential for health professionals responsible for tropical diseases control. This module is designed to that end and can be used for in-service training of professionals or as part of a disease control training course. If the latter, the module should be scheduled early on in the training course so as to serve as a foundation for the understanding of the epidemiology of specific diseases and for its application to situation analysis.

For whom is the Learner's Guide designed?

The guide is designed for health professionals with some responsibility for the control of tropical diseases. It will also give a good foundation for professionals involved in other disease control programmes. This module is not designed for the training of epidemiologists, for which several publications are available (see list at the end of this Guide).

Objectives

At the end of the training programme based on this *Learner's Guide* the learner should have acquired the skills and competence necessary and sufficient to:

1. Define and describe the role of health statistics in the implementation of an epidemiological approach to tropical diseases control

2. Define and describe major types of descriptive and analytic studies, their purpose, and their primary users

3. Define the usefulness and limitations of each type of epidemiological study in drawing conclusions about disease problems

4. Describe the steps in planning, setting up and evaluating different types of epidemiological studies

5. Examine, analyse and interpret data of epidemiological studies and surveys

6. Acquire and apply practical skills in the use of various techniques and tools for the calculation and interpretation of numerical information.

How is the course run?

Tutor and Facilitators

The tutor has considerable experience in epidemiology and statistics and can help learners resolve a wide range of problems. Facilitators are experienced professionals who work with the tutor to help the learners achieve the objectives outlined above. Facilitators will guide discussions and provide general help to individuals and to small groups of learners when necessary.

Presentations

Formal presentations of information in the guise of classroom lectures will usually be kept to a minimum and each session will be as short as possible. The information that will be given in such sessions is already contained in this Guide, so there will be very little need to take notes. A presentation will usually include examples of epidemiological principles and/or statistical calculations.

Small group work

Group work will include working through exercises and examples. A moderator chosen by the members of each group will lead discussions on particular subjects. The sessions provide good opportunities for learners to give their opinions, develop their ideas and learn from one another.

Evaluation

Evaluation of the learner

The tutor will carry out evaluation of individual progress and achievement through multiple-choice tests. Each learner will be provided with a series of questions, each one with a list of possible answers from which to select one or more answers considered as correct. The correct answers to each question will not necessarily be given at the end of these sessions, but the tutor will analyse the results to identify topics that were not clearly understood. The tutor may also identify and explain mistakes and point out areas that require improvement. This part of the evaluation is designed to help learners and tutor to assess how well the subject has been understood. Multiple-choice tests will take place at the beginning of the course and again at the end.

Evaluation of the training by the learners

By means of a questionnaire, the tutor will ask the learners how they think the training has helped them and how it might be improved. This evaluation will take place at the end of the training period in order to provide as much feedback as possible. Replies to the questionnaire may be signed or not, but the learners should feel completely free to make suggestions for improvements on the part of the tutor, as well as in the content of the course and the training facilities.

Use of the Learner's Guide

This *Learner's Guide* consists of instructional materials designed to enable the learner to achieve the objectives stated earlier. The Guide is divided into chapters called Learning Units. The skills and knowledge contained in one Unit must be acquired before progressing to the next, otherwise the learner may have difficulty in achieving the objectives of subsequent Learning Units.

A word on calculators

Recommended purchase

For most epidemiological purposes, an inexpensive calculator that can perform addition, subtraction, multiplication, division, squares, log, ln, and square roots and has a memory function will be adequate. It is also useful to have parentheses functions for the more complex calculations. For those who anticipate doing some simple statistics such as means and standard deviations, a statistical calculator that permits calculation of these two measures may be useful.

Calculators can be battery-powered or solar-powered; if it is costly or difficult to obtain batteries, the learner may prefer a calculator that is solar-powered.

How to get acquainted with a calculator?

Most calculators are supplied with an instruction booklet and give examples of actual calculations. Read through the entire booklet, but concentrate on the simple functions you are likely to be using. If your calculator contains parentheses, percents, reciprocal functions, and calculating with a constant, learn how these work since it may save the learner a great deal of time in doing complicated calculations. Work through the examples in the booklet and practise during the exercises in the following pages.

Learning Unit 1

Introduction to epidemiology

Learning objectives

By the end of this Unit, you will be able to:

- Provide a definition of *epidemiology*
- Provide a definition of *surveillance*
- Define analytic studies and describe their purpose
- Describe the major types of descriptive studies and their primary uses
- Describe the major types of analytic studies
- Provide a definition of *random error*, *bias*, *confounding* and *validity*

Epidemiology may be defined as the study of the distribution and determinants of health-related states or events (including disease) in human populations, and the application of this study to the control of diseases and other health problems. The word epidemiology consists of the Greek words επι (epi) = among, δημος (demos) = people, and λογος (logos) = doctrine.

Different methods are used in carrying out an epidemiological investigation: surveillance and descriptive studies are used to study distribution; analytic studies are used to study determinants (causes, risk factors).

Surveillance[1] and Surveys

Surveillance may be defined as the process of systematic collection, orderly consolidation, and analysis of data, with prompt dissemination and feedback of the results to those who need to know, particularly those who are in a position to take *action*. Surveillance generally uses methods distinguished by their practicality, uniformity and rapidity rather than by accuracy or completeness. It is usually based on information collected as part of routine health care, although it may sometimes be based on repeated purposive surveys.

A survey may be defined as an investigation in which information is systematically collected. The term survey is sometimes used in a narrow sense to refer specifically to a "field survey".

Uses of surveillance

- Determine trends over time.
- Gather simple information on determinants of risk or other geographical subdivisions, disease, geographical subunits or categories of individuals at risk (districts or age groups with higher rates).
- Detect the occurrence of disease (sporadic, endemic, epidemic).
- Set goals and targets based on information regarding prevalence and trends in order to design health interventions.
- Assess whether health goals and targets are being reached.

Descriptive studies

Descriptive studies may be defined as studies that describe the patterns of disease occurrence by time, place, and person.

Uses of descriptive studies

- In health planning and administration; descriptive studies and the analysis of their results allow planners and administrators to allocate resources efficiently.
- They are also used for hypothesis generation; often providing first important clues about aetiology.

[1] Document WHO/CDS/CR/ISR/99.2 *WHO recommended Surveillance Standards* covers the topic in detail for specific disorders.

Types of descriptive studies

Case reports or case series

- These describe socio-demographic, behavioural and/or medical characteristics for one or more persons with a similar diagnosis (example: characteristics of children admitted to a hospital with cerebral malaria during a one-year period).

- They provide an important link between clinical medicine and epidemiology.

- They are often useful for generating hypotheses and examining new diseases. However, conclusions about aetiology or risk factors cannot be made without having undertaken analytic studies (see below) to examine the expected frequency of exposure to the aetiological or risk factor in a group that does *not* have the illness under investigation.

Ecological Studies

- These may compare disease frequencies among different groups during the same period, or compare disease frequencies in the same population at different points in time as a function of some exposure. For instance, the increase over time in the number of persons working as gem miners along the Thai-Cambodian border (an exposure) parallels the rise in *Plasmodium falciparum* malaria cases during the same time period (an outcome).

- Ecological studies usually are quick and easy to perform, and can be undertaken with already available information, but great care is needed to avoid reaching conclusions based on spurious associations.

- Ecological studies cannot link exposure to outcome in a given individual.

- Descriptive studies constitute one of the first steps in outbreak investigation; and should always be undertaken before initiating further analytic studies.

Analytic studies

Analytic studies may be defined as studies used to test hypotheses concerning the relationship between a suspected risk factor and an outcome, and to measure the magnitude of the association effect, and its statistical significance. An analytic study always implies a comparison among two or more groups.

There are two main types of analytic studies: observational and experimental.

Observational studies

- Most analytic studies fall in this category.

- There is no human intervention involved in assigning study groups; one simply observes the relationship between exposure and disease.

- Observational studies are subject to many potential biases. Careful design and analysis may help avoid many of these biases.

- There are three basic categories of observational studies: cross-sectional, cohort studies and case-control studies

Cross-sectional Studies (Surveys)

- These examine the relationship between a disease or other health-related characteristic and other variables of interest they exist in a population at a given time. The presence or absence (or the level) of a characteristic is examined in each member of the study population or in a representative sample. These studies are used to obtain information not routinely available from surveillance or case series.

- Cross-sectional studies provide no information on the temporal sequence of cause and effect. In surveys examining the association between an exposure and an outcome, both are measured simultaneously and it is often hard to determine whether the exposure preceded the outcome or vice versa.

- Surveys may simply describe characteristics or behaviours within a study population (malaria prevalence, vaccine coverage); or may be used to examine potential risk factors (e.g., how those who receive vaccination differ from those who do not).

- In general, surveys measure the situation at a given moment – prevalence – rather than the occurrence of new events – incidence (see Unit 2).

Surveys are not recommended for the study of rare diseases or of diseases with a short duration, nor are they suited to the study of rare exposures.

Case Control Studies

Case-Control Studies proceed conceptually from outcome to exposure; start with groups affected with the outcome (in the case of a disease, the “ill” group) and groups not affected (“well”) and retrospectively determine the rates of exposure to a risk factor for each group to see if these rates differed.

- Group of subjects with the disease or other outcome variable (cases), and group of subjects without the disease or other outcome variable (controls) are identified.

Information about previous exposures is obtained for cases and controls, and the frequency of exposure is compared for the two groups.

- In case-control studies, both exposure and disease are normally considered to have occurred prior to enrolment in the study.

Cohort Studies[2]

Cohort studies proceed conceptually from exposure to outcome; starting with exposed and unexposed groups and following them to see if the rates of occurrence of the outcome in the two groups differ.

- Study groups are identified by exposure status prior to ascertainment of their disease status; both exposed and unexposed groups are then followed prospectively in an identical manner until they develop the disease under study, until the study ends, or the subjects die or are lost to follow-up. Both cohorts should have similar characteristics except for the exposure under investigation.

- Cohort studies differ from experimental studies in that the investigator does not determine exposure status. This is determined by genetics or biology (sex, presence or absence of genetic disease, etc.), subject's choice (smoking behaviour, use of contraceptives, sexual behaviours, food consumption, etc.) or other circumstances (rural versus urban, socio-economic status etc.).

- In some studies, called *retrospective cohort studies*, exposure and outcome both lie in the past (before enrolment). The main conceptual element to remember is that the retrospective cohort proceeds from exposure to disease.

[2] A cohort can be defined as a designated group of people who have had a common experience vis-à-vis exposure, and are then followed up or traced over a period of time.

Experimental studies

- The person conducting the study randomizes the subjects into exposed and unexposed groups and follows them over time to compare their rates of disease development. Examples may include trials of the efficacy of a new drug compared with the efficacy of the drug currently in use; or assessment of the efficacy of impregnated mosquito nets compared with non-impregnated nets.

- Randomization helps ensure comparability of the groups and avoids many of the biases inherent in non-experimental studies; for this reason experimental studies have been considered as a widely accepted "gold-standard".

- Experimental studies are nevertheless expensive; they are not suitable for the study of rare disease outcomes, may take a long time to perform, often present complex problems of ethics[3], or may simply not be feasible (e.g. randomized trials of the health benefits of breastfeeding). They may also provide results different from those observed under field conditions.

Potential errors in epidemiological studies

Random error

Random error is the divergence, due to chance alone, of an observation on a sample[4] from the true population value, leading to lack of precision in the measurement of an association. There are three major sources of random error: individual/biological variation, sampling error, and measurement error.

Random error can be minimized but can never be completely eliminated since we can study only a sample of the population; individual variation always occurs and no measurement is perfectly accurate. Random error can be reduced by careful measurement of exposure and outcome, thus making individual measurements as precise as possible. Sampling error occurs as part of the process of selecting study participants within a larger population. The best way to reduce sampling error is to ensure that the sample is really random and that it is of sufficient size (see Unit 8 – Surveys and data management).

[3] The ethical problems relative to epidemiological studies are addressed in *International Guidelines for the ethical Review of epidemiological Studies* issued by the Council for international Organizations of medical Sciences, Geneva , 1991 ISBN 9290360496

[4] A sample is defined as a selected subset of a population. Sampling is the process of selecting a number of subjects from all the subjects of a population.

Systematic error (bias)

Bias occurs when there is a tendency to produce results that differ in a systematic manner from the true values. A study with small systematic bias is said to have high accuracy. Bias (or systematic error) may lead to over- or underestimation of the strength of an association.[5] The sources of bias in epidemiology are many and over 30 specific types of bias have been identified. The main biases are:

- Selection bias
- Information bias
- Bias due to confounding.

Selection bias
Selection bias occurs when there is a systematic difference between the characteristics of the people selected for a study, or who agree to participate, and the characteristics of those who are not selected, or who do not agree to participate (e.g. in a study limited to volunteers).

*Information bias (*also called *observation bias)*
Information bias occurs when there are quality (accuracy) problems in the collection, recording, coding or analysis of data among comparison groups. Interviewers might, for example, interview the cases with more diligence than the interview control, or a person with a disease may recall previous exposures better than persons who are healthy (this type of bias is called *recall bias).*

Although selection bias or information bias can usually be corrected at the time of analysis, it is best to think about possible sources of bias at the time of the study design so that they can be minimized or avoided.

Bias due to confounding
In a study of the association between exposure to a cause (or risk factor or protecting factor) and the occurrence of the disease, confounding can occur when another factor exists in the study population and is associated both with the disease and the initial factor being studied. A problem arises if this second extraneous factor is unequally distributed among the exposure subgroups. Confounding occurs when the effects of two protective or risk factors have not been separated and it is therefore incorrectly concluded that the effect is due to one variable rather than the other. For instance, in a study of the association between tobacco smoking and lung cancer, age would be a confounding factor if the average ages of the non-smoking and smoking groups in the study population were very different, since lung cancer incidence increases with age.

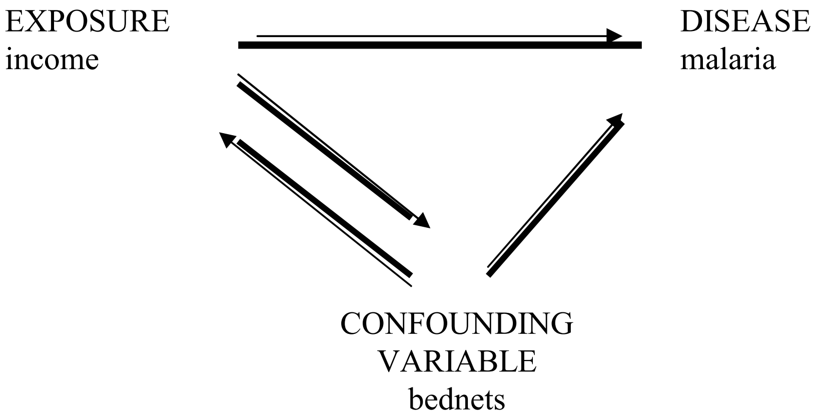
Another example of confounding is shown in Figure 1a. Suppose one wishes to study the relationship between income and malaria, illustrated by the top line of Figure 1a. It

[5] Statistical dependence among two or more variables, which are said to be associated if they occur together more frequently than would be expected by chance. Statistical tests permit calculation of the association.

is likely that a higher income protects against malaria (e.g. people with a high income are more likely to buy drugs for malaria prophylaxis). It is also known that, in that community, income is associated both with the use of bednets. The relationship between income and malaria is thus affected by the relationship between bednets and income. In other words, bednets confound the relationship between income and malaria.

Such biases can be controlled for in the analysis if appropriate information has been collected during the study on potential confounding variables[6] and if each factor is properly analysed and interpreted.

Figure 1a: Confounding – bednets, income and malaria



Validity and Reliability

Validity is an expression of the degree to which a test is capable of measuring what is intended to be measured. A study is valid if its results correspond to the truth; there should be no systematic error, and the random error should be as small as possible. Figure 1b illustrates the relationship between the true value and measured values for low and high validity and reliability (or repeatability). With low reliability but high validity the measured values are spread out, but the average of the measured values is close to the true value. High reliability does not ensure validity since the measurements may all be far from the true value.

[6] The very term "malaria" relates to a confounding factor, since it relates to "bad airs", once thought to be a factor for the disease – although this not a direct causal factor.

Figure 1b: Validity and reliability

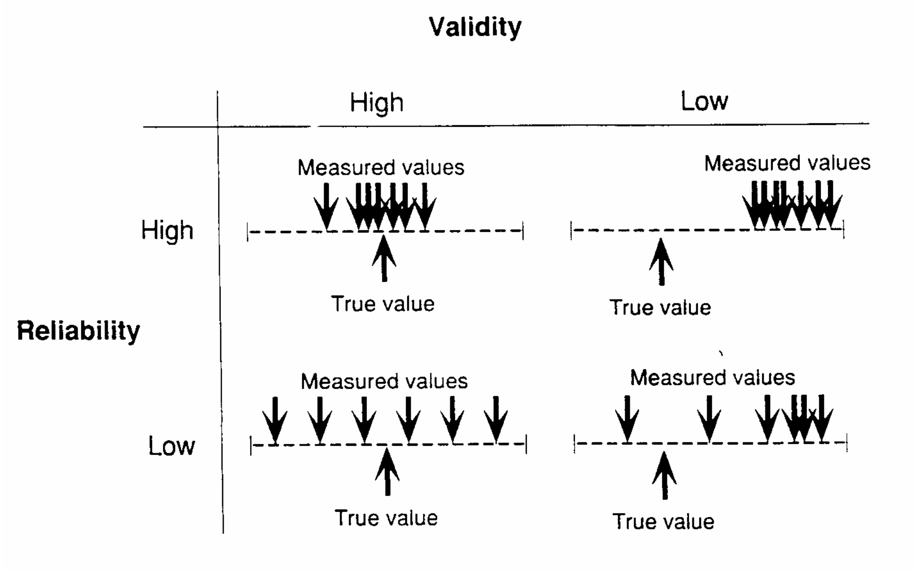
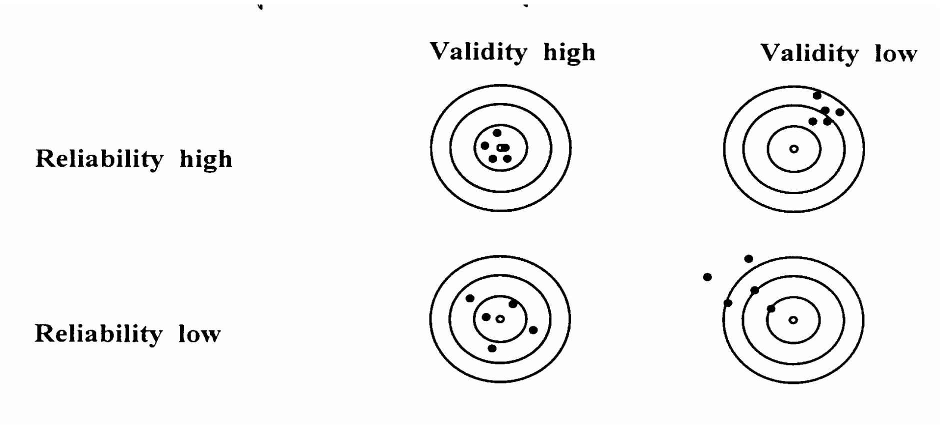


Figure 1c shows the same concept in a different graphical way: the centre of the target corresponds to the true population value; individual target shots are individually measured from 5 samples in each example.

Figure 1c: Validity and reliability

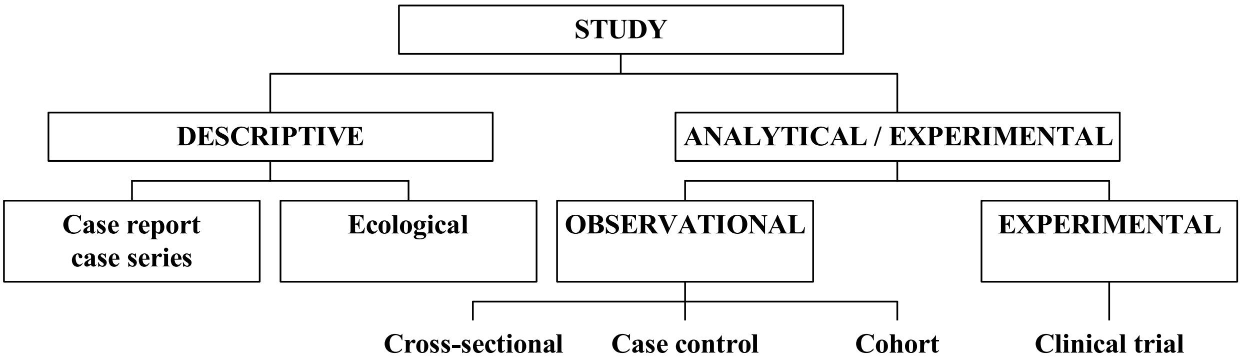


The logical sequence of epidemiological studies

In epidemiological research, the current state of knowledge often determines the most logical study design. One usually sees a progression from hypothesis-generating to hypothesis-testing studies. For example, hypotheses are often generated by methods such as surveillance, case reports, case series, or ecological studies. These hypotheses are then tested using data from experience, from previous cross-sectional studies, from case-control studies, or from retrospective cohort studies, which can be done relatively quickly and cheaply. If these studies lend support to the hypothesis, a prospective cohort study may be undertaken. Finally, in some situations, a randomized clinical trial may be undertaken.

The following flowchart illustrates the application of the various types of studies In all types of study, hypothesis setting must precede analysis.

Figure 1d: Types of studies



Learning Unit 2

Rates, ratios and proportions

Learning objectives

By the end of this Unit, you will be able to:

- Define the terms *rate*, *ratio* and *proportion*
- Differentiate incidence rate from prevalence rate and give examples of their uses
- Differentiate point prevalence from period prevalence
- Determine the correct denominator for the calculation of each of these terms
- Calculate rates, ratios, and proportions using appropriate numerators, denominators and constants
- Apply the concepts of rate ratios and rate differences
- Apply the concepts of standardization of rates

Depending on who is using the data and for what purposes, the data may be presented as **raw data**, **proportions**, **rates** and **ratios**.

Raw data

Raw data may be defined as the entire set of data collected in a study, before any rounding, editing or statistical organization. They are of use primarily in helping health planners and administrators determine health care needs. A health planner may want to know the **number** of cases of malaria seen in the last year so that they can estimate how many chloroquine tablets to order for the next year. He or she may also want to know approximately how many deliveries take place each year so as to staff the obstetrics services appropriately.

Any variable can be considered as one of two types: discrete or continuous.

Discrete variables have values that can fall into only a limited number of categories without intermediate levels (e.g., gender – M/F, survival –dead or alive, exposure status –Yes/No, race, marital status…). When the possible categories have a natural order of progression, the variables are called *ordinal* (e.g., improvement in mobility or level of current cigarette smoking – none, light, moderate, heavy). Some quantitative data can also be discrete e.g., parity (it is not possible for a woman to experience a fraction of a live birth).

Continuous variables can assume all possible values along a continuum within a specified range (e.g., height, weight, blood pressure…). They are limited only by the accuracy and precision of measurement.

Proportion

This is also a measure that is of use primarily to planners and administrators. It is defined as the percentage of the total number of events which occur in a data set, usually expressed as a percentage. The formula is ***(x/y)k***, where **x** is the number of individuals or events in a category and **y** is the total number of events or individuals in the data set and ***k*** is a constant, in this case 100.

Example: *Of the 120 cases of malaria admitted to hospital X last year, 80 were children. The proportion (percentage) of children among cases is (80/120) x 100 or 66.7%.*

It may be useful for the hospital administrator to know that 67% (two-thirds) of malaria hospitalizations occur in the paediatric age group and 33% (one-third) occur in adults. He or she can thus plan the number of beds and staffing of various categories required to take care of malaria patients. However, for public health, these data are of limited use, since they say nothing of who is at risk for getting malaria.

Rate

For the public health practitioner interested in determining who is at risk and monitoring the success of prevention efforts, the most useful measure is a rate. Rates measure the relative frequency of cases in a population *during a specified period of time*. The general formula is the same as for proportions, namely ***(x/y)k***, although here ***x***, ***y***, and ***k*** take on different meanings. Rates may measure incidence (new cases) or prevalence (newly occurring plus pre-existing cases) within a specified period.

An **incidence rate** is the occurrence of new cases of a disease within a defined population at risk during a specified period of time. In this situation:

- *x* is the number of new cases in the defined population which had its onset during a specified period of time
- *y* is the average size of the defined population at risk in which the disease could occur during the time period specified, usually at the middle of the time period

- *k,* a constant, depends on convention or is the value such that the smallest rate in the data set has at least one digit to the left of the decimal point.

An **attack rate** is a variant of an incidence rate. In practice, the attack rate will only differ from the incidence rate if there is a large proportion of persons in the population who are not at risk (for instance, children who have been successfully vaccinated against measles may be considered not to be at risk for the disease).

In a **prevalence rate**, *x* is the number of existing cases, new and old, in a defined population during a specified period (period prevalence) or at a given point in time (point prevalence).

Example: *In January, 3 new cases of trachoma were detected in a village. There were already 10 people in the village who had the disease, but two successfully completed a course of therapy during the month and were considered cured. The population of the village was 2600. In this case:*

- *the incidence rate is (3/2600) x 1000 or 1.2 per 1000 or 0.1%*[7]
- *the period prevalence rate is (3+10)/2600) x 1000 or 5 per 1000 or 0.5%*
- *the point prevalence rate as of 31 January is ((3+10-2)/2600) x 1000 or 4.2 per 1000 or 0.4%.*

Ratio

A **ratio** is an expression of the relative frequency of the occurrence of some event compared to some other event, for example, the ratio of cases among males to cases among females. Here, the formula is also ***(x/y)k,*** where:

- *x* is the number of events or persons having a specified attribute
- *y* is the number of events or persons having an attribute different from those of the event or person in x
- *k* is 1.

In this situation, the ratio is often expressed as *x:y*, with *y* usually equal to 1 (*y* can be made equal to 1 by dividing both *x* and *y* by *y*).

Example: *If there are 15 male cases (x) and 5 female cases (y) of a given disease, the male:female ratio can be calculated as 3:1 by dividing both values by 5 (y).*

Ratios are often used when it is difficult to ascertain the population denominator for a disease or a condition correctly. One example is the abortion ratio, which is the

[7] Strictly speaking, the attack rate is the number of cases occurring during January in the population at risk (excluding those already affected), i.e., [3/(2600-10) * 1000] = 0.1 per thousand, equal to the incidence rate. The Incidence Rate = 3/(2600 total – (10 affected –2 cured and therefore sensitive again) i.e. 3/2592. In practice, these requirements are often neglected where they make little difference.

number of abortions divided by the number of live births during the same time period. The formula remains *(x/y)k* with *k* determined either by convention or by the value that gives at least one digit to the left of the decimal point.

Relative risk and risk difference

Rates for two or more groups (males/females, age categories, educational levels, presence or absence of some behaviour) are often compared by dividing one by the other or by subtracting one from the other.

If they are divided, the result is called a **rate ratio** or **relative risk**. The formula is simply rate *a/*rate *b*, where *a* is the incidence in the group exposed to the factor under investigation and *b* the incidence rate in the group not thus exposed. The **rate ratio** or **relative risk** may be used to identify possible causal risk factors and identify markers that may be useful in targeting services. A ratio of one means that there is no difference in outcome between the exposed and the unexposed groups (if the outcome is an incidence rate, this will be the same for both exposed and unexposed groups). A ratio over one suggests that the characteristic (exposure) is a risk factor; a ratio of less than one suggests a protective effect.

Example: *People who go into the forest have a malaria incidence rate of 10/1000 per month, while people who do not go into the forest have a malaria incidence rate of 1/1000 per month. The risk ratio is (10/1000)/(1/1000) or 10. Thus, people who go into the forest are 10 times more likely to get malaria than those who do not.*

Example : *People who use nets have a malaria incidence rate of 2/1000 per month; people who do not use nets have a rate of 8/1000 for the same period. The ratio of the rates is (2/1000)/(8/1000) = 0.25. Thus, those who use nets incur a lower rate of malaria incidence than those who do not (this is called the protective effect and is calculated as 1—the relative risk or 1-0.25 = 0.75 – this is roughly equivalent to saying that 75% of those who use bednets in these circumstances will be protected against malaria).*

Example: *People who are illiterate have a malaria incidence rate of 8/1000, while those who are literate have a rate of 4/1000 for the same period. The ratio of the rates is 2. Thus, those who are illiterate have twice the rate of malaria than those who are not.*

Here, literacy is a marker rather than a causal risk factor. Illiteracy does not cause malaria, but those who are illiterate are at risk for other reasons, such as living conditions, occupation, etc.

Rates may also be compared by subtracting one from the other. The resulting value is known as the **(absolute) risk difference**. This is calculated as **rate** *a* minus **rate** *b*. This represents the absolute differences in risk between the exposed and the unexposed groups. The absolute difference is less suitable than the relative risk in indicating a causal effect. If disease incidence is the same in the exposed and in the unexposed group, the value of the absolute risk difference will be zero. If there is a causal relationship between the characteristics being studied and the outcome, the risk difference provides information on the amount of disease that could be prevented if the characteristic could be eliminated.

Example: *Those going into the forest have a malaria incidence rate of 10/1000 per month; those who do not have a malaria incidence rate of 1/1000 per month. The risk difference is 10/1000 - 1/1000 or 9/1000. The absolute difference between the groups is 9 per 1000. Because there is a presumably causal relationship, you could say that if people stopped going into the forest, the malaria rate would be reduced by as much as 9/1000, to 1/1000.*

Caution is required in making such statements since often people will have more than one characteristic (exposure) that puts them at risk for a disease; eliminating only one behaviour or characteristic usually does not fully solve the problem. Caution is required in making such statements if the characteristic is a marker rather than a causative factor; changing a marker without changing the causal factors associated with it is unlikely to result in a lower disease rate. As in the example above, reducing illiteracy will not *per se* reduce the incidence of malaria, although it may lead to socio-economic changes that will. Other criteria such as biological plausibility, laboratory evidence, a dose-response effect, and the fact that exposure precedes outcome must also be taken into account before drawing conclusions about causality.

Standardization of rates

Standardization is a process permitting comparisons among sets that show different compositions for confounding factors (e.g. age and sex). For instance, an age-standardized rate is used to eliminate the effect of age differences in the populations compared. An age-sex standardization rate is used to eliminate the effect of differences in the age and sex distribution.

Take for instance the overall malaria mortality in each of two sets hereafter. This will depend in part on the age composition of each set.

	Region A	**Region B**
Total population	76 521	61383
Deaths	701	1001
Crude death rate per 1000	**9.16**	**16.30**

Is mortality in A less than mortality in B?

Standardization can be

- direct
- indirect

Direct method of standardization

The crude death rate in A is lower than that in B. However, if we look at the age-specific death rates in A and in B in Table 1, we note that the death rate is *higher* in region A for all age-groups except age 60-69.

Table 1 Mid-year population by age and age-specific death rates for A and B

	Region A			Region B		
Age	**Population**	**Deaths**	**Deaths/1000**	**Population**	**Deaths**	**Deaths/1000**
0-4	9222	250	27.11	6473	156	24.10
5-14	19576	44	2.25	13740	27	1.97
15-49	39056	117	3.00	22458	65	2.89
50-59	4156	50	12.03	6400	74	11.52
60-69	2688	80	29.76	6140	198	32.17
70-79	1489	114	76.56	5200	374	72.00
80+	334	46	137.72	972	107	110.08
	76521	**701**	**9.16**	**61383**	**1001**	**16.30**

Let us consider the age-specific rates for a third (standard) population C which in this case is the sum of both populations (A and B) with rates that are averaged between the two populations (e.g. for age group 0-4 years, the standard population is 9222+6473 = 15695 and the death rate is (27.11+24.10)/2 = 25.61) – population and rate for each of the remaining age-groups of population C are calculated in the same way.

Table 2 Mid-year population by age and age-specific death rates for three regions

	Region A		Region B		*Region C (standard)*	
Age	**Population**	**Deaths/1000**	**Population**	**Deaths/1000**	***Population***	***Deaths/1000***
0-4	9222	27.11	6473	24.10	*15695*	*25.61*
5-14	19576	2.25	13740	1.97	*33316*	*2.11*
15-49	39056	3.00	22458	2.89	*61514*	*2.95*
50-59	4156	12.03	6400	11.52	*10556*	*11.78*
60-69	2688	29.76	6140	32.17	*8828*	*30.97*
70-79	1489	76.56	5200	72.00	*6689*	*74.28*
80+	334	137.72	972	110.08	*1306*	*123.90*
	76521	**9.16**	**61383**	**16.30**	***137904***	***12.40***

Steps in the computation of standard death rates for the two regions by the direct method.

For each region, calculate the number of deaths expected in each age-group and in the total population if the *death rate* of that region is applied to the corresponding *population* numbers for region C. For instance, the death rate for age-group 0-4 in region A is 27.11/1000—applied to the corresponding population of 15 695 in region C, this leads to an expected number of deaths of 27.11 * 15 695 = 425 deaths.

You similarly calculate the number of deaths for each successive age-group and add up the total (1767) for region A. Do the same for region B (total number of expected deaths = 1653)

Table 3 Applying rates of A and B to population C

Age	*Standard*	Region A		Region B	
	Population C	Deaths/1000	Expected deaths (Rate A * standard pop. C)	Deaths/1000	Expected deaths (Rate B * standard pop. C)
0-4	*15695*	27.11	425	24.1	378
5-14	*33316*	2.25	75	1.97	66
15-49	*61514*	3.00	185	2.89	178
50-59	*10556*	12.03	127	11.52	122
60-69	*8828*	29.76	263	32.17	284
70-79	*6689*	76.56	512	72	482
80+	*1306*	137.72	180	110.08	144
Total	***137904***	**9.16**	**1767**	**16.30**	**1653**

The standardized death rate (SDR) is :

$$SDR = \frac{\text{Total number of expected deaths x 1000}}{\text{Standard Population}}$$

For Region A $SDR\ (A) = \frac{1767 \text{ x } 1000}{137904}$ **= 12.81**

For Region B $SDR\ (B) = \frac{1653 \text{ x } 1000}{137904}$ **= 11.98**

Interpretation

Although the crude death rate (CDR) is higher in region B (16.30 per 1000) than in region A (9.16 per 1000), the standardized rate is lower for region B (11.98 per 1000) than for region A (12.81 per 1000). The proportion of the population aged 60 and more is far higher in region B (6140 + 5200 + 972 / 61383 = 0.20) than in region A (2688 + 1489 + 334 / 76 521 = 0.06) and the higher crude death rate in region B is therefore not surprising. When the populations of the two regions are standardized by age and compared on the basis of similar age composition, the effects of age on the crude death rates are eliminated. It may be concluded that the mortality risks in the population of region A are higher than those in region B.

Indirect method of standardization

In a similar way, we can calculate an expected number of deaths for regions A and B by applying the age-specific *rate in standard population C* for each age-group to the *populations of A and B* in that age-group, respectively.

Steps in the computation of standard death rates for the two regions by the indirect method.

Age	***Rate of deaths/1000 for standard population C***	**Population A**	**Population B**
0-4	*25.61*	9222	6473
5-14	*2.11*	19576	13740
15-49	*2.95*	39056	22458
50-59	*11.78*	4156	6400
60-69	*30.97*	2688	6140
70-79	*74.28*	1489	5200
80+	*123.90*	334	972
	12.40	**76521**	**61383**

For each region, calculate the number of deaths expected in each age-group and in the total population if the *standard* death rate is applied to the corresponding population numbers for that region.

For instance, the population for age-group 0-4 in region A is 9222. Applying the 0-4 standard death rate in C (26.00/1000) to a population of 9222 leads to an expected number of deaths of:
26.00 * 9222/1000 = 240 deaths. You similarly calculate the number of deaths for each successive age-group and add up the total (654) for region A. Do the same for region B (total number of expected deaths = 953).

Table 4 Applying rates of C to populations A and B: Expected death A and B

Age	***Deaths/1000 for pop C***	**Population A**	**Expected Deaths (Rate C * pop. A)**	**Population B**	**Expected Deaths (Rate C * pop. B)**
0-4	*25.61*	9222	236	6473	166
5-14	*2.11*	19576	41	13740	29
15-49	*2.95*	39056	115	22458	66
50-59	*11.78*	4156	49	6400	75
60-69	*30.97*	2688	83	6140	190
70-79	*74.28*	1489	111	5200	386
80+	*123.90*	334	41	972	120
	12.40	**76521**	**677**	**61383**	**1033**

$$\text{Index death rate} = \frac{\text{Total number of expected deaths*1000}}{\text{Population}}$$

$$\text{For Region A} = \frac{\text{total number of expected deaths for A*1000}}{\text{total population for A}} = \frac{677*1000}{76\ 521} = 8.84$$

$$\text{For Region B} = \frac{\text{total number of expected deaths for B*1000}}{\text{total population for B}} = \frac{1033*1000}{61\ 383} = 16.83$$

For each region, divide the crude death rate of the standardized population (12.40) by the index death rate of the region to obtain the standardizing factor.

A: Crude death rate of standard population / Index death rate A = 12.40/8.84 = 1.40
B: Crude death rate of standard population / Index death rate B = 10.50/15.52 = 0.74

Multiply the crude death rates for regions A and B by the respective standardizing factor to obtain the standardized death rates for each region:

A: Crude death rate for A * standardizing factor for A = 9.16 * 1.40 = 12.83
B: Crude death rate for B * standardizing factor for B = 16.83 * 0.74 = 12.01

Interpretation

Both the direct and indirect methods of standardization allow to take into account the effect of age. The crude death rates of region A and B are 9.16 and 16.30 respectively.

Standardized death rates (direct method) are 12.81 for region A and 11.98 for region B; these results are similar to those obtained through indirect standardization (12.83 and 12.01).

When age structure is taken into account, overall mortality risk is seen to be higher in region A than in Region B.

A note on rounding

The procedure for finding the last digit of a measure is called “rounding”. There are three general rules for rounding:

Rule 1: if the digit *beyond the last digit* to be reported is less than 5, drop everything after the last digit to be reported. Rounding to one decimal place, the number 5.3467 becomes 5.3.

Rule 2: if the digit after the last digit to be reported is greater than 5, increase the last digit to be reported by one. The number 5.798 becomes 5.8 when rounding to one digit.

Rule 3: to prevent rounding bias, if the last significant digit is exactly 5, it is general practice to round to the integer preceding the 5, and rounding up if this is an odd integer.

Thus the number 3.55 (rounded to one digit) would be 3.6 (rounding up) and the number 6.450 would round to 6.4 (rounding down when rounding to one decimal).

It is also possible to round by taking the nearest whole number: 66.7% may be rounded to 67%.

Exercises: Rates, ratios and proportions

Exercise 1

The following table presents malaria morbidity data for Province X in Africa, which has received a large number of immigrants in recent years:

Number of malaria cases, province X, 1990-1994

YEAR	CASES	POPULATION
1990	30 858	492 810
1991	36 602	585 540
1992	46 172	738 870
1993	56 439	891 280
1994	68 392	1 044 620

a) Describe in words the trend in the number of cases.

b) Calculate the incidence rate of malaria cases/100 population for each year and describe the trend in words.

c) Compare the trend in the number of cases and the trend in rates. How do you explain your observations?

d) Which is the more appropriate measure to monitor changes over time in the area?

Exercise 2

The adjoining province Z (population 169 250) had 15 233 cases in 1994.

a) Which province had the higher rate in 1994?

b) In your opinion, which area should receive the greatest funding for control efforts and why?

Exercise 3

A survey among children in Region A shows that 450 out of 950 children have evidence of parasites in their blood.

a) What is the parasite rate?

b) Is this an incidence rate or a prevalence rate? Why?

Exercise 4

In 1994, 49 140 malaria cases occurred among males and the remaining 23 250 occurred among females

a) What is the ratio of male:female cases?

b) What percentage of the total cases occurred in males? In females?

Exercise 5

At one of the health centres in province X, the age breakdown of malaria cases was as follows:

Age breakdown of malaria cases, Province X, 1995

AGE	CASES	% OF ALL CASES	POPULATION AT RISK	INCIDENCE RATE/100
0-11 MONTHS	71	2.4	1980	3.6
1-4 YEARS	645	21.9	7920	8.1
5-14 YEARS	698	23.7	12 300	5.7
≥15 YEARS	1528	51.9	27 300	5.6
TOTAL	**2942**	**100.0**	**49 500**	(5.1)

a) Which age group accounts for the biggest percentage of all cases?

b) Which age group is at greatest risk of contracting malaria?

c) Why are the answers to a and b different?

Exercise 6

A study shows that the incidence rate of malaria is 10/1000 cases per week among Thai villagers who work as gem miners and go into the forests, whereas this rate is 2/1000 cases per week among farmers from the same villages.

a) Calculate the relative risk of malaria among gem miners

b) Interpret your findings in words

c) Calculate the risk difference between the gem miners and the farmers

d) Interpret your findings in words.

Learning Unit 3

Data presentation: tables, graphs and charts

Learning objectives

By the end of this Unit, you will be able to:

- List the features of good tables, graphs, and charts
- Plot and label a series of tables, graphs and charts correctly from raw data
- List the uses for semi-logarithmic presentation.

Tables

A table may be defined as a set of data arranged in rows and columns designed to present the frequency with which some event occurs in different categories or subdivisions of a variable, as can be seen from Table 3.1 on the next page.

Guidelines for developing tables

- Keep them simple. Better 2 or 3 small tables than a single large table.
- No more than 3 variables should be used in a table.
- All tables should be self-explanatory:
  - Clear and concise title telling **what**, **where**, and **when**
  - Rows and columns must be clearly labelled
  - Units of measurement must be stated
  - Codes, abbreviations, and symbols must be footnoted
  - Totals must be shown
- If data are not original, their source must be footnoted

Table 3.1

Proportion of malaria cases in relation to total in-patients, Kulak State.
1990 - 1994

rows and columns to be clearly labelled

clear and concise title telling what, where and when

Year	All patients	Malaria patients	%
1990	136 289	16 946	12.4
1991	114 327	18 117	15.8
1992	101 050	13 821	13.7
1993	79 485	10 757	13.5
1994	76 403	11 533	15.1

Data collected from 24 district hospitals.

footnote to enhance clarity

Graphs

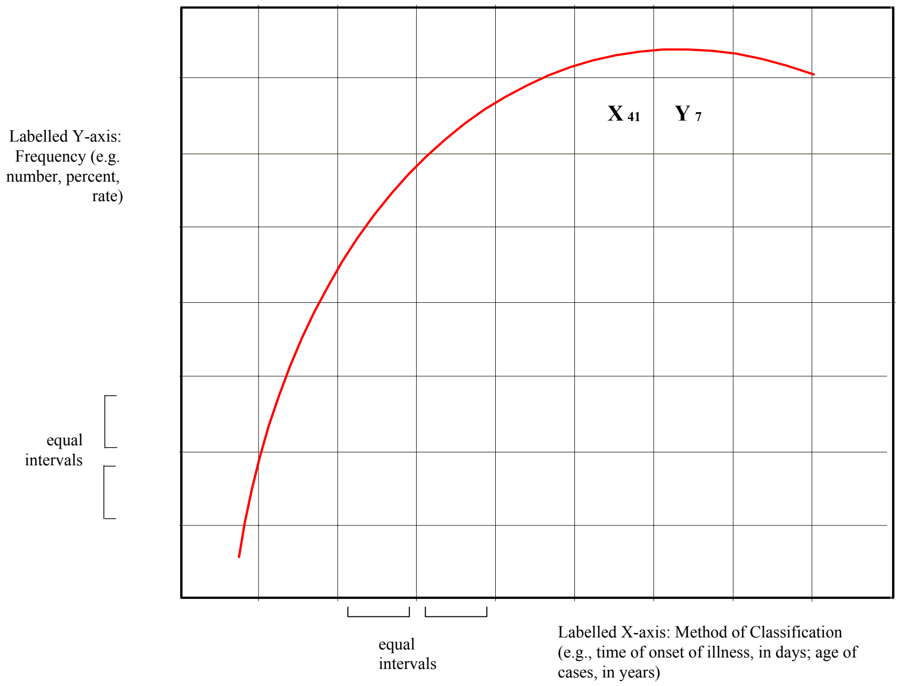
A graph may be defined as a method of showing quantitative data using a drawing on a coordinate system. The most common form is a rectangular coordinate, with two sets of lines at right angles to each other and divided into equal intervals. The *x* axis by convention is the horizontal axis, and the *y* axis is the vertical one.

Graphs are used for continuous variables such as time, parasite counts etc. Charts rather than graphs are used for non-continuous variables such as sex or educational level.

The variable (age, year, etc.) is usually classified along the *x* axis; the *y* axis is the axis generally used for measures of frequency. See Figure 3a.

Figure 3a

General graph with descriptive annotation



Guidelines for graphs

- Keep them simple and do not try to put in too much information.
- Every graph should be self-explanatory:
- Avoid interrupting the axis (scale breaks) if at all possible.
- Title.
- Axes clearly labelled.
- Units on the x and y axis clearly specified.
- Equal quantities must be represented by equal intervals on an axis; on the *x* axis, categories covering 10 years, for example, should be twice as long as categories covering 5 years.

Suggestions for graphing data

- Start by placing the data in a simple table.
- Examine the range of values on the *x* and *y* axes.
- Count the squares in the horizontal and vertical direction on your graph paper and decide which way to put the graph on the paper.
- Divide the number of squares on the *x* axis by the number of units (i.e. number of months or years) needed in order to determine how many spaces will be represented by each unit; repeat for the *y* axis. Round the calculated value down to the nearest whole number.

Example: Suppose your graph paper has 14 large horizontal squares and 20 large vertical squares, each of which is divided into 5 smaller squares. You are plotting the monthly number of cases of malaria for the past four years, and the number of cases ranges from 120 to 450 each month. In this case, you may decide to place the months (x axis) on the long side of the paper and the number of cases (y axis) along the short side. To determine how many squares that one month should correspond to, take the total number of small squares (20x5) and divide by the number of months (48). The result is 2.1, which should be rounded down to 2. Thus, each month should be represented by two small squares.

Similarly, along the y axis the range of values is 450. In this case, you would take the number of small squares (14x5) and divide by 450. In this case, the answer is 0.16. This should be rounded down to 0.1. Thus, each case corresponds to 0.1 of the small squares, or 10 cases correspond to 1 small square, and each large square will correspond to 50 cases.

- Label the squares on the *x* and *y* axes at appropriate intervals.

Types of graphs

Remember: your graph must have enough information to be clear without further explanations

Charts

The most common forms are bar charts, pie charts and geographic coordinate charts. Applications are given here for 620 patients, classified according to 4 age categories (<1, 1-5, 6-10 and 11-15). In the <1 group there were 250 patients; in the 1-5 group there were 325; in the 6-10 group there were 30; and in the 11-15 group 15 patients.

Pie Charts

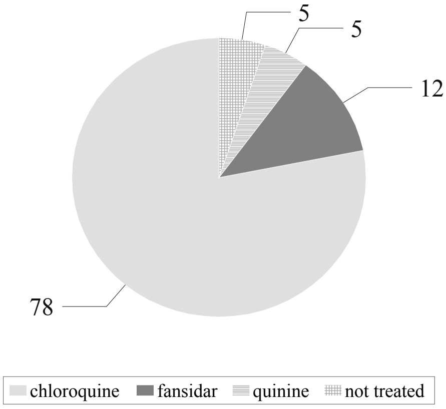
This is defined as a circular chart most frequently used to show percentage distributions that uses wedge-shaped portions proportionate to the size of the category. The convention is to start at 12:00 and arrange "slices" anticlockwise in order of decreasing size. To convert from percentages to degrees, multiply the percentage by 3.6.

Example 60% when converted into degree is:
$60 \times 3.6 = 216^{o}$ or
$0.60 \times 360 = 216^{o}$

See also Figure 3b.

Figure 3b

Treatment of paediatric malaria patients prior to hospitalization
Lenin Hospital
Zanzibar Town January - March 1989



Bar Charts

Bar charts have cells, all of which have the some column width whatever the size of the category. The bars may be arranged vertically or horizontally. By convention, there is always a space between the bars. Bar charts are easier to use when categories are of unequal size; they *must* be used if categories are not continuous (i.e. sex, marital status, etc.), as is the case in Figure 3d. Figure 3c is a bar chart of continuous data (age) with categories of unequal importance.

Figure 3c

Age distribution of hospitalized paediatric patients Lenin Hospital, Zanzibar Town, January - March 1988

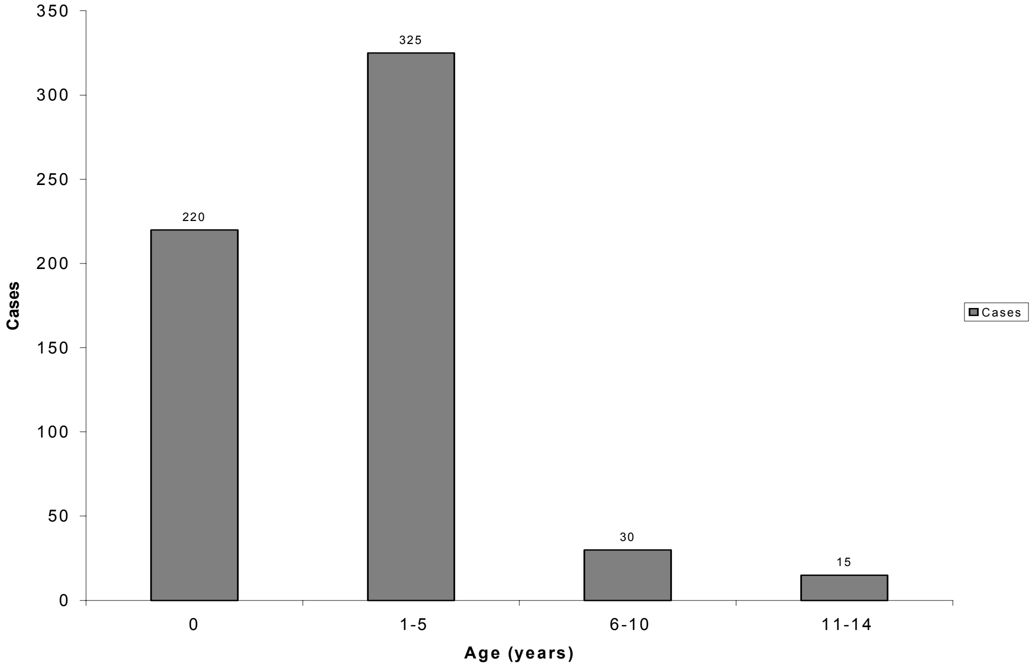
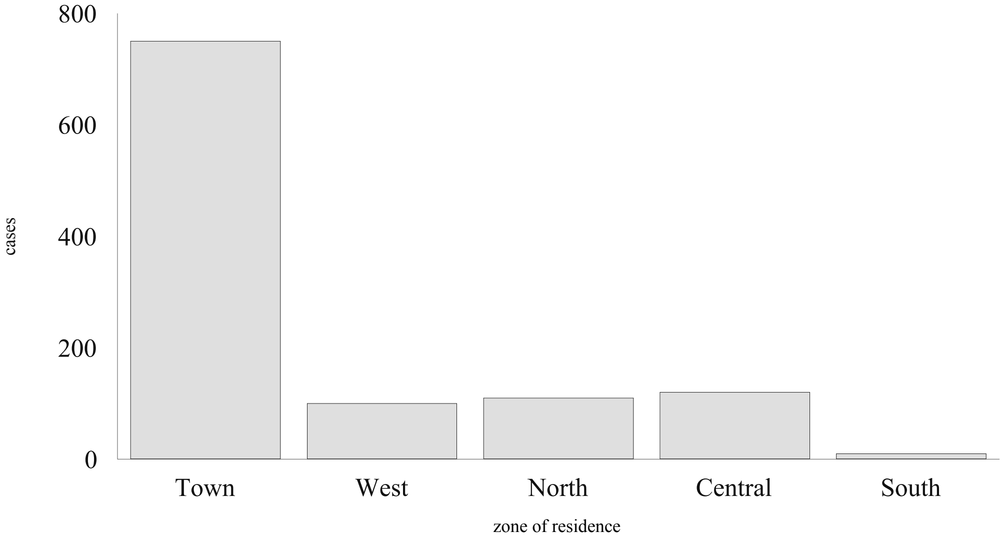


Figure 3d

Cases of malaria in hospitalized paediatric patients
Lenin Hospital, Zanzibar Town
January - March 1988

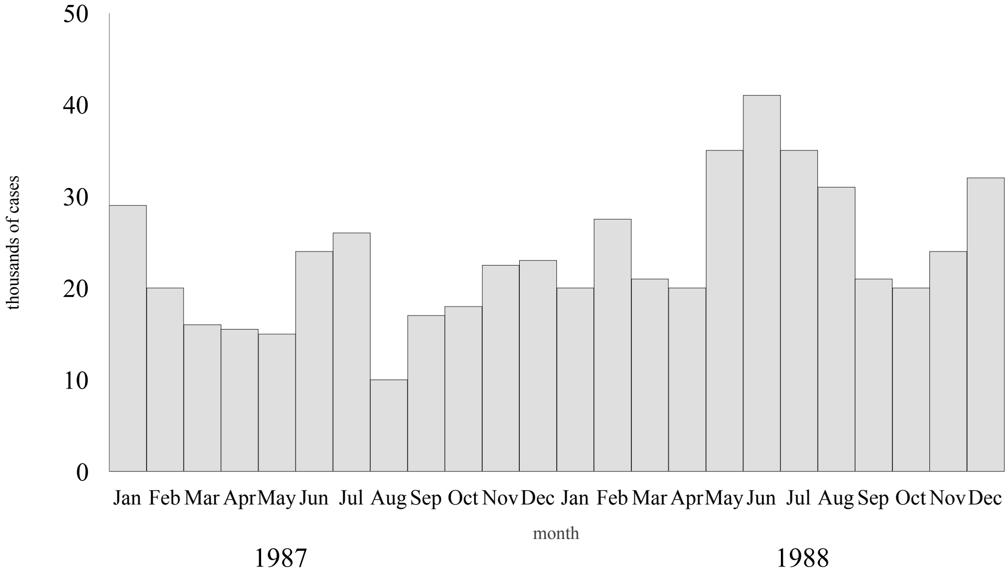


Histogram

This may be defined as a bar graph of the frequency distribution of a **continuous** quantitative variable in which the width of the bar is proportional to the unit of value of the variable on the *x* axis and the height of the bar is proportional to the unit of value of frequency on the *y* axis. By convention, there is no space between the bars, and no scale breaks are allowed on the *y* axis.

Figure 3e

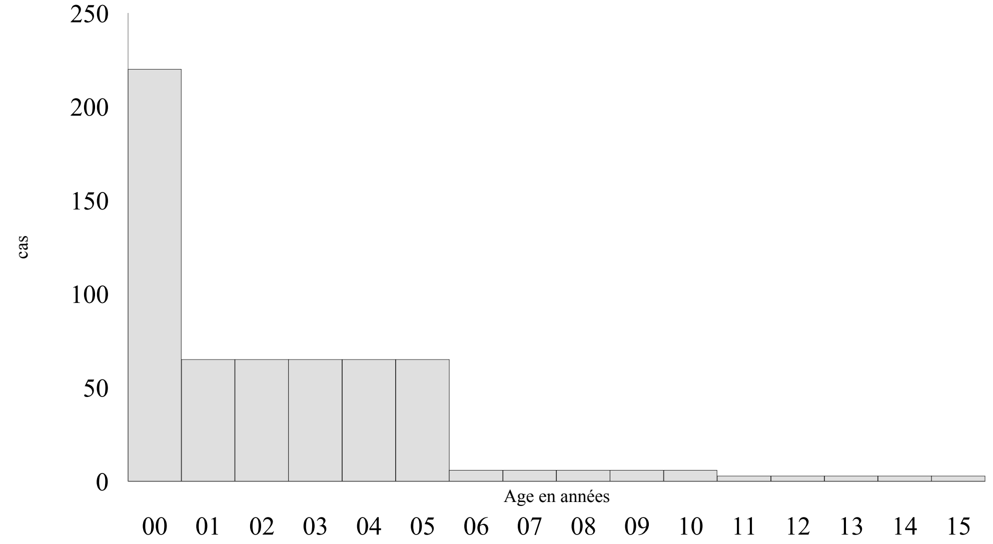
Monthly distribution of malaria cases reported by PHC units and health centres, Zanzibar State 1987-88



Histograms can be used to plot the number of cases or percentages on the *y* axis, but are generally not used to plot rates. Additionally, although subgroups of the cases are occasionally plotted within the bars (example: deaths among the cases), some authors prefer using a frequency polygon.

N.B: If categories are of different sizes, the height of the bar must be adjusted so that the area contained within the bar is proportional to the number of individuals within the category. See Figure 3f; in this situation there were 4 age categories (<1, 1-5, 6-10 and 11-15). In the <1 group there were 220 patients; in the 1-5 group there were 325; in the 6-10 group there were 30; and in the 11-15 group 15 patients. To represent these figures correctly as a histogram, the height of the bar is determined by the total number of patients in the category divided by the number of units in the category. Thus, in the 0-1 year age group, the height of the bar is equal to the number of cases. In the 1-5 year age group, the height of the bar for each year of age in that category is calculated by dividing the total cases 1-5 years by 5: 325/5 = 65. The height of the column for each year in the 6-10 year age-group is similarly 30/5 = 6, and 15/5 = 3 for the 11-15 year age-group.

Figure 3f
Age distribution of hospitalized malaria paediatric patients
Zanzibar January-May 1989



Line graphs

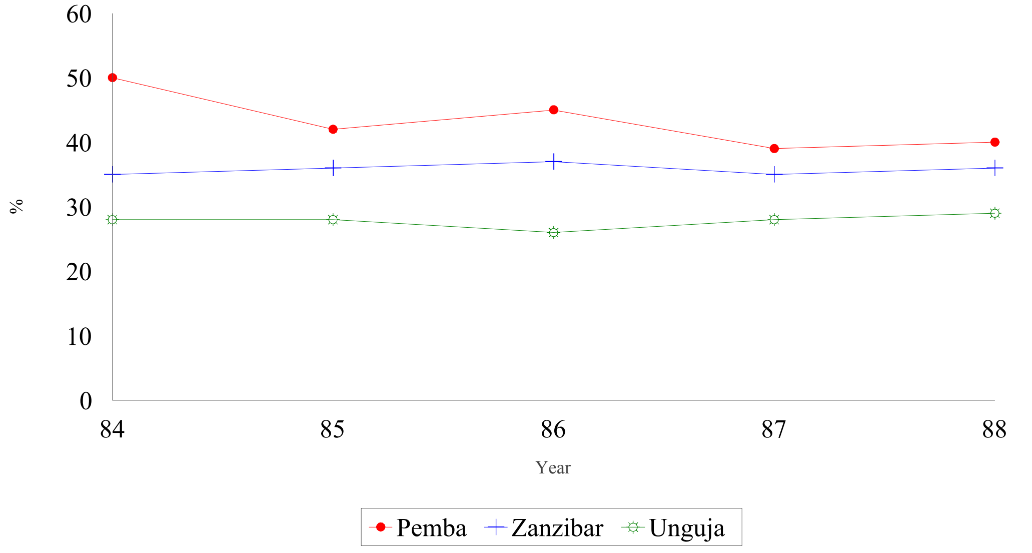
This may be defined as a graph of the frequency distribution of a continuous variable created by plotting the frequency of a category on the *y* axis at the midpoint of the category on the *x* axis. Values for each category are connected by a continuous line.

If a graph is to contain the frequency distribution by category for more than one group (i.e. the frequency of cases over a ten-year period for males and females), it may be advisable to use line graphs.

Line graphs may be used to plot number of cases and percentages; they are the method of choice for plotting rates. See Figures 3g and 3h.

Figure 3g

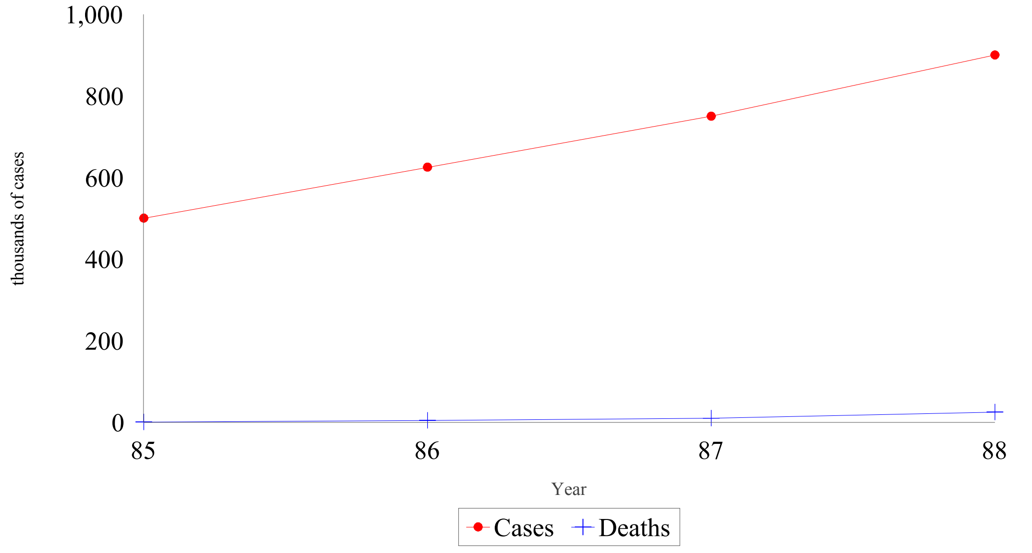
Rate of positivity for malaria blood films examined at hospitals level, Zanzibar State*
1984 - 1988



*by Island

Figure 3h

Malaria cases and deaths officially notified in health facilities run by physicians, Madagascar
1985 - 1988



Geographic coordinate charts

This is a chart where cases are marked as dots on a map according to the number of cases, or areas shaded geographically according to the incidence or prevalence rate of the disease considered. See Figures 3i and 3j hereafter.

Figure 3i
Map: Health centre, houses, trees, road, river, *Biomphalaria* (schistosomiasis) and *Simulium* (onchocerciasis)

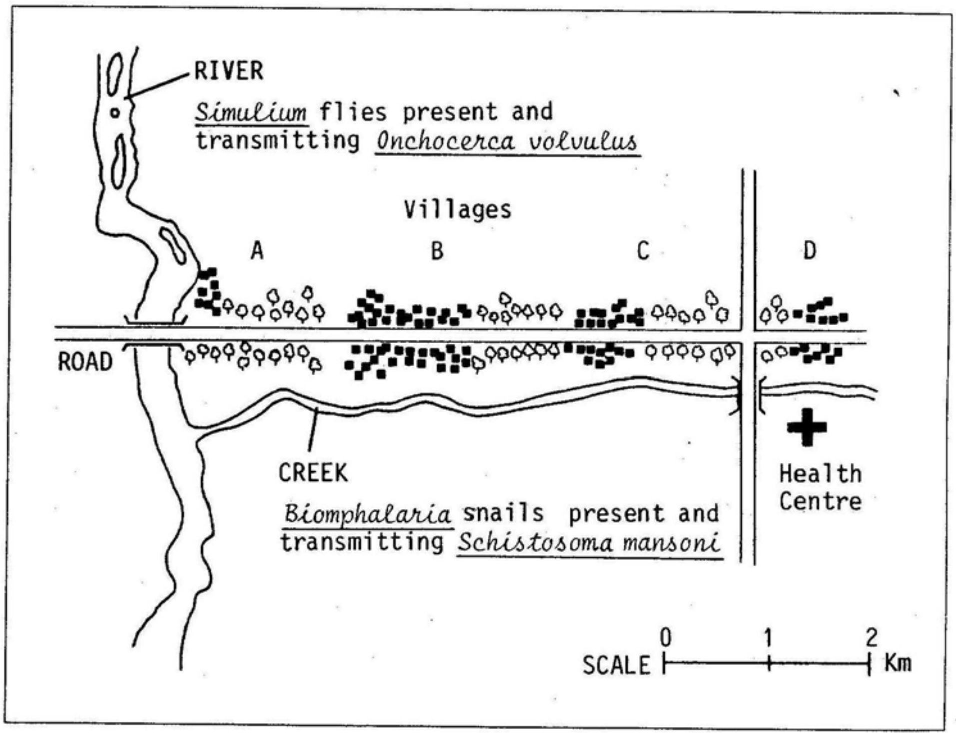
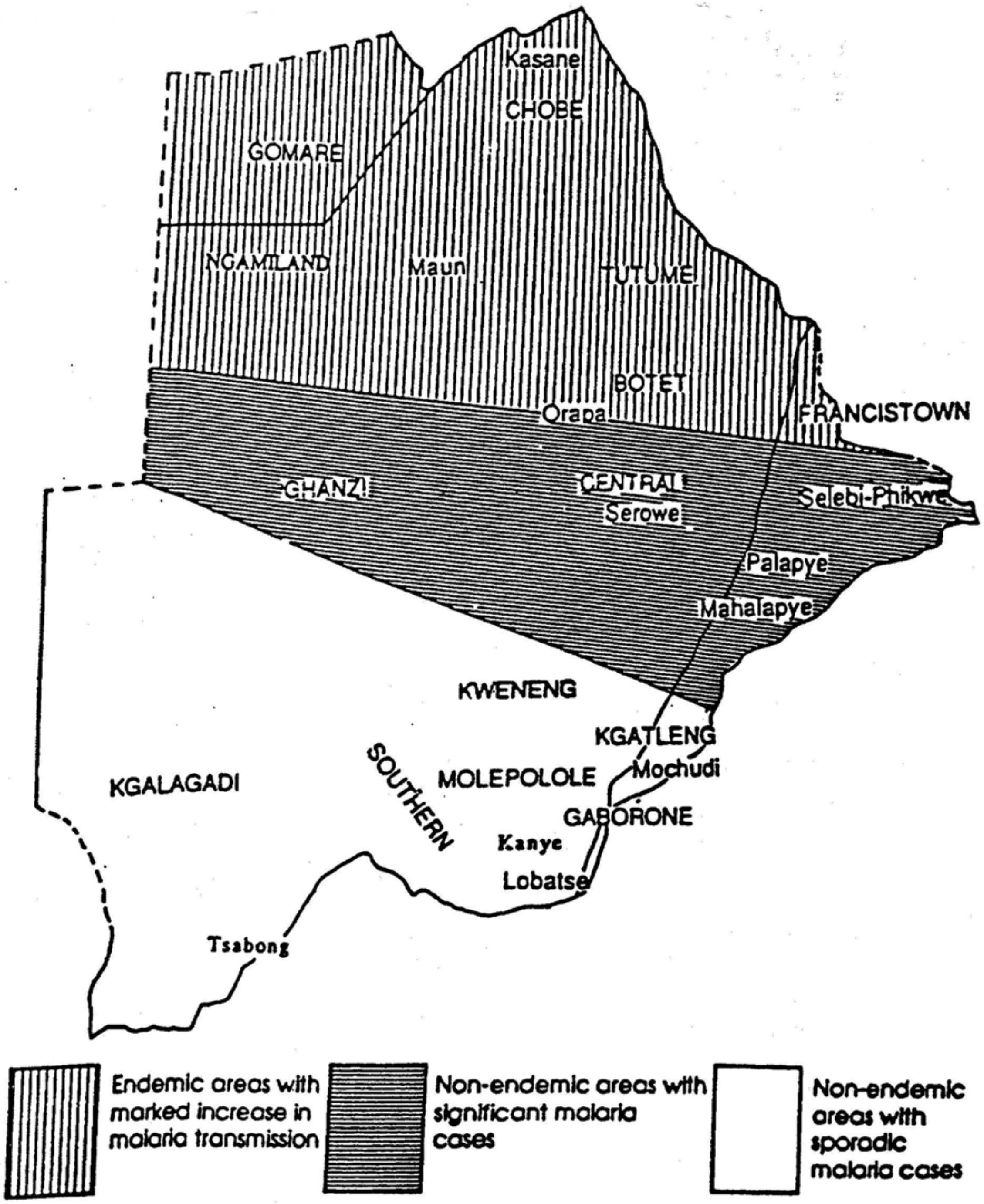

RIVER
Simulium flies present and transmitting Onchocerca volvulus
Villages
A
B
C
D
ROAD
CREEK
Biomphalaria snails present and transmitting Schistosoma mansoni
Health Centre
SCALE
0
1
2
Km

Figure 3j: Distribution of malaria, 1996 (Botswana)



Semi-logarithmic graphs

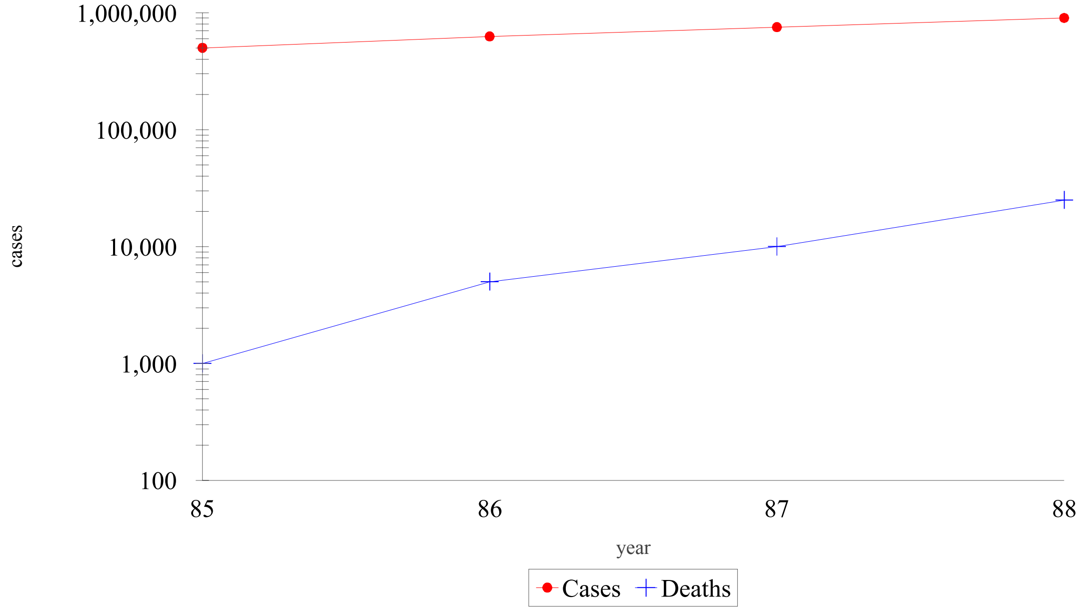
This may be defined as a graph in which the y axis is measured in logarithms of units and x axis is measured in arithmetic units. These graphs are generally used to:

- Plot data when the range is too great to present meaningfully on an arithmetic graph.
- Examine relative rather than absolute changes over time.

If a line plotted on a semi-logarithmic graph is straight, it indicates a constant rate of change, and the slope allows direct measurement of the rate of change. Two or more lines that follow parallel paths have equal rates of change. See Figure 3k, which represents a logarithmic plot of the data shown in Figure 3j and allows a better look at the trends and relative rates of change for cases and deaths.

Figure 3k

Malaria cases and deaths officially notified in health facilities run by physicians
Madagascar 1985 - 1988



Note: With the generalization of computer graphics software programmes, charts, maps, graphs etc are increasingly done on the computer; the use of semi-log paper is on the wane.

Exercises: Data presentation

The Government of a large Asian country started a national programme 5 years ago to reduce the morbidity and mortality from 3 major childhood diseases. Each District Medical Office (DMO) is now being asked to evaluate the effectiveness of this programme in their district and to learn as much as possible about who remains at risk for developing these diseases. The DMO of Mantu has decided that the best way to begin is to use the past five years of surveillance data to examine the trends over the past five years in the incidence of these three diseases. The surveillance data provide the number of cases and deaths by month for each disease and are broken down by broad age categories, including one for children under 5.

Exercise 1

What are the advantages and disadvantages of using surveillance data to monitor trends for the 3 diseases? What other sources of data might the DMO consider to gather information on trends in the 3 diseases?

Exercise 2

The DMO has the number of cases and deaths for children under 5 for each disease. What other number(s) does the DMO need, to monitor disease trends over a several year period in an adequate manner? Where can the DMO obtain such numbers?

Exercise 3

The most recent national census was conducted in 1987, and there are no population estimates available for the years 1988-1992. The population in 1987 in the under 5 age group was 56 650. The rate of natural increase for the population is 3.3% per year.

How can the DMO estimate the under 5 population in the district for each of the years 1988 to 1992?

Exercise 4

After estimating the mid-year population of children under 5 for each year between 1988 and 1992, the DMO develops a table containing data for the three diseases over the five-year period (see Table 1 below).

Divide your group into 3 smaller subgroups. Each subgroup should perform one of the following tasks:

(a) Plot the trends for the incidence of each of the 3 diseases on the same graph.

(b) Plot the trends for mortality from each of the 3 diseases on the same graph.

(c) Plot the trends for case fatality rates in each of the 3 diseases on the same graph.

Describe the trends you have graphed to the remainder of your group.

Exercise 5

Each subgroup will perform one of the following tasks:

(a) For the disease assigned to your group, plot the age distribution from a hospital record review (Table 2).

(c) Plot the seasonal distribution based on 5 years of surveillance data for the disease assigned to your group (Table 3).

Exercise 6

Taking into account the graphs you have prepared, define the characteristics of the disease assigned to your group. Describe in words the trends for incidence, mortality and case-fatality as well as the age and seasonal distribution of the disease, and what types of actions or events may have been responsible for the temporal trends observed. Prepare a summary to present to the rest of the class. Presentations will be limited to 10 minutes per group.

TABLE 1 (exercise)
Incidence, Mortality, and Case Fatality Rates for Diseases A, B, and C,

Mantu District, 1988-1992

Disease A

Year	<5 population	<5 cases	<5 deaths	Cases/1000	Deaths/1000	Case-fatality %
1988	58 520	10 241	205	175	3.5	2.0
1989	60 541	10 353	157	171	2.6	1.5
1990	62 446	10 616	131	170	2.1	1.2
1991	64 507	10 966	123	170	1.9	1.1
1992	66 635	11 261	113	169	1.7	1.0

Disease B

Year	<5 population	<5 cases	<5 deaths	Cases/1000	Deaths/1000	Case-fatality %
1988	58 520	3113	152	53.2	2.6	4.9
1989	60 541	1604	85	26.5	1.4	5.3
1990	62 446	4571	219	73.2	3.5	4.8
1991	64 507	1251	71	19.4	1.1	5.7
1992	66 635	2259	113	33.9	1.7	5.0

Disease C

Year	<5 population	<5 cases	<5 deaths	Cases/1000	Deaths/1000	Case-fatality %
1988	58 520	480	386	8.2	6.6	80.5
1989	60 541	454	394	7.5	6.5	86.7
1990	62 446	381	356	6.1	5.7	93.4
1991	64 507	348	329	5.4	5.1	94.4
1992	66 635	347	320	5.2	4.8	92.3

TABLE 2 (exercise)
Age Distribution for Diseases A, B and C
Mantu Distinct Hospital 1 January - 31 December 1992

Number of cases

Age group months	A	B	C
0-5	427	37	258
6-11	1063	296	10
12-23	2312	411	3
24-35	1239	223	24
36-59	647	115	35
Total	**5688**	**1082**	**330**

TABLE 3 (exercise)
Monthly Distribution
for diseases A, B and C 1988-1992

Percent of cases

Month	A	B	C
January	3	13	8
February	7	10	7
March	15	9	6
April	18	7	8
May	13	5	9
June	41	4	7
July	7	4	8
August	5	5	9
September	3	7	12
October	6	9	9
November	7	12	8
December	5	15	9
Total	**100**	**100**	**100**

Learning Unit 4

Measures of central tendency

Learning objectives

By the end of this Unit, you will be able to:

- Define the terms *mean*, *median*, and *mode*
- Describe the advantages and disadvantages of using the mean versus the median
- Calculate means, medians and modes from individual and from grouped data.

With variables such as age, number of children, haemoglobin, and parasite counts, it is often useful to develop a single value that is representative of the individual values in the group. These single, representative values are known as measures of central tendency. These values not only facilitate the description of a population, but also facilitate the comparison of populations.

The most common measures of central tendency are the **mean,** the **median**, and the **mode**. These measures can be calculated from individual data if the number of items in the data set is small; if there are many items, these measures are instead calculated from grouped data.

The mean

The mean is the **average** of the values in the data set. It is calculated by taking the sum of the individual values in the data set and dividing this sum by the number of values in the set. The mean is the most commonly used measure of central tendency, in part because it is used in other, more sophisticated statistical tests. Its major disadvantage is that it can be affected by the presence in the set of a few extreme values, large or small.

Mathematically, the formula can be expressed as follows:

$\overline{X} = \square\ x_i/n$ where:

$\overline{X}$ is the arithmetic mean

$\square$ is the sum of

X_i is each individual value in the set

n is the number of individual values in the set.

Example: *Calculate the mean of a set of 5 values: 12, 15, 7, 13, 8*

$$\overline{\mathrm{X}} = \frac{12+15+7+13+8}{5} = \frac{55}{5} = 11$$

The median

The median is the value or point in a data set that divides the ranked values into 2 equal-sized groups, one consisting of values smaller than the median, and the other consisting of values greater than the median. If the data set is skewed in one direction or another (with one or more extreme values), the median is a more representative measure of central tendency than the mean, because it does not take outliers into account.

The median is calculated as follows:

- Order the values by rank (place the values in sequence, either in ascending or in descending order)

- Identify the mid-point of the sequence; if there is an **odd** number of values identify the middle number, if there is an **even** number of values, identify the mid-point between the two numbers in the middle of the sequence.

The general formula to identify the middle value is:

$$\text{Middle value} = \frac{\text{total number of values in sequence} + 1}{2}$$

- The number corresponding to this middle value is the median of the values in the data set.

Example: *Determine the median of a data set containing an odd number of values (12, 15, 7, 11, 8).*

- *Rank values in ascending order: 7, 8, 11, 12, 15*
- *Determine the mid-point of the sequence (5 values +1) / 2=3.*
- *The median is therefore the third value in the sequence*
- *The third value is 11, therefore the median is 11.*

Example: *Determine the median of a data set with an even number of values where the middle numbers are different (12, 15, 18, 7, 13, 8).*

- *Rank values in ascending order: 7, 8, 12, 13, 15, 18*
- *Determine the midpoint of the sequence (6 values+1)/2=3.5.*
- *The median is the value that lies halfway between the 3d and 4th values.*
- *The 3d and 4th values are 12 and 13. The median is (12+13)/2 = 12.5.*

Example: *Determine the median of a data set with an even number of values where the middle numbers are the same (12, 15, 18, 7, 12, 8).*

- *Rank values in ascending order: 7, 8, 12, 12, 15, 18*
- *Determine the midpoint of the sequence (6 values+1)/2=3.5.*
- *The median is the value that lies halfway between the 3d and 4th values.*
- *The 3d and 4th values are 12 and 12. The median is (12+12)/2 = 12.*

The mode

The mode is the value in a set of data, which occurs most frequently. It is identified by counting the number of times a value occurs in the data set and determining which one occurs most often. Sometimes a data set may have more than one mode.

Example: *The mode of the values 12, 15, 18, 7, 12, 8, 3, 19, 2 is 12 because this number appears twice while the others appear only once.*

Example: *The sequence of values 12, 15, 12, 3, 18, 7, 12, 8, 3, 15, 19, 3, 2, has two modes, 3 and 12, since both numbers appear 3 times while the others occur only once or twice.*

Calculation of the mean, median and mode for grouped data

If there are many observations in a data set, or if the observations have been grouped together for other purposes, the mean, median, and mode can be calculated using the frequency distributions.

Example:
Haemoglobin levels in grammes among Amazonian Gold Miners, Brazil 1990

Hb in GRAMMES	NUMBER OF CASES
8.0 - 8.9	2
9.0 - 9.9	4
10.0 - 10.9	9
11.0 - 11.9	14
12.0 - 12.9	7
13.0 - 13.9	2
14.0 - 14.9	1
15.0 - 15.9	0
16.0 - 16.9	2
	n = 41

Mean

In order to calculate the mean, the formula is:

$\overline{X} = \Sigma f_i x_i / n$ where

$\overline{X}$	is the arithmetic mean
Σ	is "the sum of"
f_i	is the frequency of occurrence of an event (e.g. the numbers in column 2 in the above table)
x_i	is the measurement units or the midpoints of the intervals (see column 3 below)
n	is the number of individual values in the data set.

In order to determine the mean haemoglobin value, it is necessary to add a third and fourth column to the above table:

Haemoglobin levels in grammes among Amazonian Gold Miners, Brazil, 1990

1	2	3	4	5
CLASS INTERVAL Hb in grammes	f_{ii} Number of cases	Cumulative number of cases	x_{Ii} Midpoint of interval shown	$f_I * x_I$ Product of columns 2 & 4
8.0 - 9.9	6	6	8.95	53.70
10.0 - 11.9	23	29	10.95	251.85
12.0 - 13.9	9	38	12.95	116.55
14.0 - 15.9	1	39	14.95	14.95
16.0 - 17.9	2	41	16.95	33.90
TOTAL	41			**470.95**

Applying the formula, we now have

$\Sigma f_i x_i = 470.95$ and $n = 41$, hence

$\overline{X} = \Sigma f_i x_i/n = 470.5$ grammes/41 = 11.5 grammes

Median

As with the individual data, the first step in calculating a median is to determine the middle case, which is (total number of cases + 1)/2. Here, there are 41 cases and the middle case is the 21st case (class interval 10.0-11.9). We now must find which haemoglobin value corresponds to the 21st observation. This is accomplished by calculating the cumulative number of cases, starting with the lowest haemoglobin value:

The 21st case is included in the range 10.0-11.9 (cases 17-29). Next we apply the following formula to determine the exact median value:

Median = $L + (JW/f)$

where:

L true lower limit of the class interval containing the median point (in the present case, $L = 10$)

J number of cases in this interval below the midpoint case, calculated as the number of cases below the midpoint (here = 21) minus the cumulative number of cases up to (but not including) this interval (here = 6), which is 15

W width of the class interval = 2

f number of cases in this class interval (in this case $f = 23$) .

Applying the formula, we get:

Median = $L + (JW/f)$ = 10 + (15 x 2/23) grammes/millilitre = 11.30 grammes/millilitre.

Mode

The mode is the category in which the greatest number of cases are seen. In this case it is 10.0-11.9 since this interval has the greatest number of cases (23 cases).

Exercises: Measures of central tendency

Exercise 1

Last month, 20 patients with malaria were admitted to hospital X. The age distribution is as follows (in years):

4, 3, 3, 1, 2, 26, 64, 3, 2, 5, 7, 4, 22, 3, 1, 1, 12, 2, 3, 6

a. What is the mode?
b. What is the median?
c. What is the mean?
d. Why are there differences between the median and the mean?
e. Which is a better measure of the age distribution in the population, and why?

Exercise 2

The duration of hospitalization for the 11 children admitted last month to hospital X with cerebral malaria was as follows:

Child 1	03 days
Child 2	07 days
Child 3	08 days
Child 4	05 days
Child 5	04 days
Child 6	11 days
Child 7	06 days
Child 8	10 days
Child 9	05 days
Child 10	01 day
Child 11	04 days

a. What is the mode?
b. What is the median?
c. What is the mean?
d. Why are the median and the mean closer than in exercise 1?
e. Which is a better measure of the distribution of length of stay of the population in this case? Why?

Exercise 3

The following data on parasite density were obtained for 200 consecutive patients seen in clinic X during the first quarter of 1991:

Density: parasites/1000 WBC	**Frequency**
1000 – 2999	20
3000 – 4999	70
5000 – 6999	80
7000 – 8999	25
9000 – 10999	5
Total	**200**

Calculate:

a. The mean parasite density

b. The median parasite density

c. The modal parasite density.

Learning Unit 5

Measures of variability and chi-squared test of association; normal distribution

Learning objectives

By the end of this Unit, you will be able to:

- Define the terms *range*, *standard deviation* and *normal distribution*
- Describe the advantages and disadvantages of the above
- Calculate a range and a standard deviation
- Calculate and interpret a chi-squared value.

The measures of central tendency (mean, median and mode) are very useful for describing a frequency distribution, but they do not indicate the spread of values that may have the same central tendency. In making decisions for the management of tropical diseases, as in many other public health fields, it is important to establish what is "normal". The "normal" value is a statistical concept and depends, to a great extent, on the distribution of the attribute in the population. The extent of variability can be summarized through 2 measures.

- The range
- The standard deviation.

The range

The range indicates the distance between the highest and the lowest value in the distribution.

Example: the range of 11 values 3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10 is 3 to 10. Range can also be expressed as 10 –3=7.

The range is simple to calculate and easy to understand. However, the range tells only about two values of a series of observations. An extremely high or low value may be due to a measurement error. The range does not take into account variability of observations between the two extreme values.

The standard deviation

The standard deviation is a measure that describes the scatter of observations around their mean. If all the observations had the same value, the standard deviation would be zero; the farther apart from one another (and from the mean) the individual observations are, the larger the standard deviation is. If the standard deviation of a sample is very small, the sample average closely represents every individual value; a large standard deviation indicates this is not so.

The steps in the calculation of the standard deviation are as follows:

- calculate the difference between each observation and the mean $(x_i - \bar{x})$
- square each difference $(x_i - \bar{x})^2$
- add the above squares and divide this sum of squares by the number of observation minus 1, i.e., (*n -1*);
- calculate the standard deviation by finding the square root of the number obtained in the above steps.

in application of the formula

$$SD = \sqrt{\frac{\sum_i^n \left(x_i - \bar{\mathrm{x}}\right)^2}{\mathrm{n} - 1}}$$

where:

x_I	each value
$\bar{x}$	the mean
$(\mathrm{x}_\mathrm{i} - \bar{x})^2$	the square of each difference
$\sum$	"the sum of"
n	number of observations.

You will note that the denominator is *n - 1* rather than *n*. In practice, when *n* is reasonably large, it makes little difference which is used. However, for theoretical reasons, *n - 1* is preferred.

Example: calculate the standard deviation (SD) of a set of 11 values: 3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10.

Following the above steps:

- calculate the mean: 66 / 11 = 6

- calculate the difference between the value of each observation and the mean:

Mean	6										
Values x_I	3	4	4	5	6	6	6	7	7	8	10
Difference from mean x_I- 6	-3	-2	-2	-1	0	0	0	1	1	2	4
Squared difference from mean	9	4	4	1	0	0	0	1	1	4	16

- add the squares of differences

  sum of squares = 9 + 4 + 4 + 1 + 0 + 0 + 0 + 1 + 1 + 4 + 16 = 40

- divide by number of observations minus 1 *(n-1)*, where $n = 11 = \frac{40}{10} = 4$

- calculate the standard deviation by finding the square root of the result:

SD = $\sqrt{4} = 2$

A quicker calculation can be:

$$SD = \frac{\sum x^2 - (\sum x)^2 / n}{n-1}$$

Where: $\sum x^2$= take each observation, square it, then sum the squares.

$(\sum x)^2$= sum the observations, then square the sum.

Using the values of the previous example:

3, 4, 4, 5, 6, 6, 6, 7, 7, 8, 10
(n = 11)

- If we square each observation and then sum the squares, we have:

  9 + 16 + 16 + 25 + 36 + 36 + 36 + 49 + 49 + 64 + 100 = 436.

  thus $\sum x^2 = 436$

- If we then sum the observations and square the sum, we have:

$(3 + 4 + 4 + 5 + 6 + 6 + 6 + 7 + 7 + 8 + 10)^2 = (66)^2 = 4356$

thus $(\sum x)^2 = 4356$

with

$$\frac{(\Sigma x^2)}{n} = \frac{4356}{11} = 396$$

and

$$SD = \sqrt{\frac{(436 - 396)}{10}} = \sqrt{4} = 2$$

The standard deviation is based on all observations, therefore it is better suited than the range for describing the distribution.

Standard deviation of a percentage

If a sample (of at least 30 subjects) yields a percentage of p%, it is possible to calculate the standard deviation of this percentage in the population as follows:

$$\mathrm{SD} = \sqrt{\frac{\mathrm{p(100 - p)}}{n}}$$

Consider a sample of 100 persons of whom 80% women and 20% men. The standard deviation of the percentage of women in the population is:

$\mathrm{SD} = \sqrt{\frac{\mathrm{p(100 - p)}}{n}}$ with p = 80, 100-p = 20, n= 100.

Here $\mathrm{SD} = \sqrt{\frac{\mathrm{p(100 - p)}}{n}} = 4$

We would expect less than five chances in one hundred that the proportion of women in one sample is less than
80% minus (1.96 * 4) or 80% - 7.84% or 72.16%
– or more than
80% plus (1.96 * 4) or 80% + 7.84% or 87.84%

Note: if you are not using a calculator, you may approximate 1.96 by 2 to simplify calculations.

Test of association: the Chi-Squared (χ2) test

The chi-squared (χ^2) test is commonly used to examine the null hypothesis that the distributions of categorical variables are independent of each other (i.e. that the frequency falling into a particular category of variable A is the same for all categories of variable B).

The table hereafter shows the distribution of two variables A and B and how to calculate the chi-squared statistic to test for an association.

*NB: The example is given for a 2 * 2 distribution with (2-1)(2-1) = 1 degree of freedom, for which the threshold values of chi-squared statistic is 3.84 at 5% probability. For degrees of freedom greater than 1, see table attached.*

Observed values

Variable B	Variable A			
		Present	Absent	**Total**
	Present	A	B	**a + b**
	Absent	C	D	**c + d**
	Total	**a + c**	**b + d**	**a + b + c + d**

Establish the null hypothesis and calculate the *expected* frequencies (E) for each observed (O) cell under the null hypothesis of independence (no association). If the null hypothesis were not rejected, one would have expected values as follows:

Expected values

Variable B	Variable A			
		Present	Absent	**Total**
	Present	(a+c) (a+b) / N	(b+d) (a+b) / N	**A + B**
	Absent	(a+c) (c+d) / N	(b+d) (c+d) / N	**C + D**
	Total	**a + c**	**b + d**	**N = a + b + c + d**

Determine the degree of freedom of the distribution (the freedom to choose frequencies in the cells under the constraint of fixed marginal totals); this is equal to
(number of columns of data minus 1) * (number of columns of data minus 1)

Calculate the chi-squared statistic

chi^2 = $\sum \frac{(O-E)^2}{E}$ with a degree of freedom (df) = (r-1)(c-1)

Compare the result with the theoretical distribution of chi-squared to determine significance. If the calculated chi^{2} is greater than the tabulated value the null hypothesis can be rejected at the corresponding level (5%, 10%) of significance. For 1 degree of freedom, the value of chi-squared corresponding to p = 0.05 is 3.84.

For instance:

Association between recent meat consumption & enteritis necrotica, Papua New Guinea

Observed values

Disease	Exposure			
		Present	Absent	**Total**
	Present	50	11	**61**
	Absent	16	41	**57**
	Total	**66**	**52**	**118**

- Null hypothesis: absence of association between food and disease
  The *expected* value for the first cell (exposure and disease both +) would be

Expected values

Disease	Exposure			
		Present	Absent	**Total**
	Present	*66 * 61 / 118 = 40.16*		**61**
	Absent			**57**
	Total	**66**	**52**	**118**

- Number of degrees of freedom = 1
  The table can then be filled by difference.

Expected values

Disease	Exposure			
		Present	Absent	**Total**
	Present	*66*61/118 = 40.16*	*61-40.16 = 20.84*	**61**
	Absent	*66-40.16 = 25.84*	*57-25.84 = 31.16*	**57**
	Total	**66**	**52**	**118**

The absolute value of the difference between *expected* and observed value is (50 – 40.16) = 9.84 for all cells; the square of this difference is 96.82.

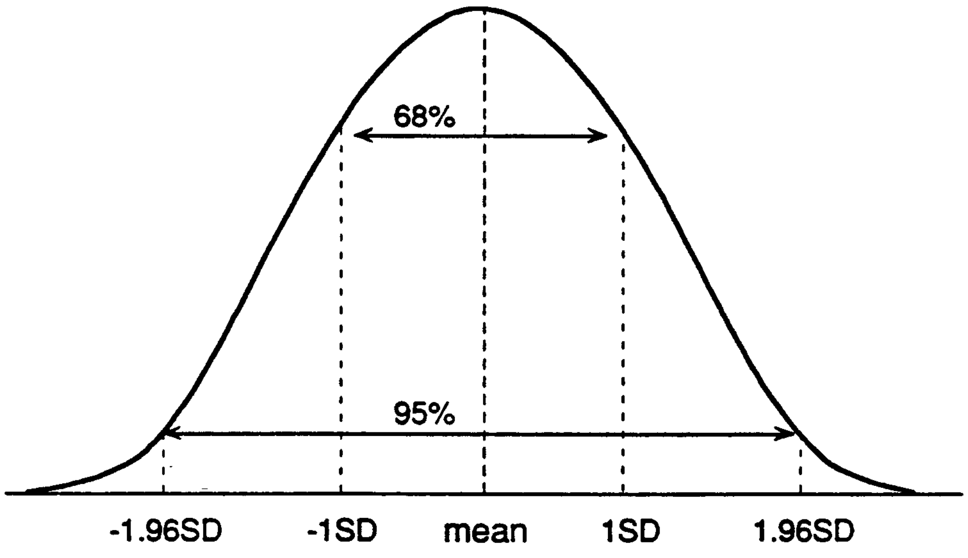
Chi-squared is the sum of (squared difference between *Expected* and Observed values/Observed value) for each cell =
96.82/40.16 + 96.82/25.84 + 96.82/20.84 + 96.82/31.16 = 2.41 + 4.65 + 3.75 + 3.1 = 13.91
This is greater than 3.84; we can reject the null hypothesis and accept the existence of an association between consumption of meat and enteritis necrotica.

NB: In practice, it is often sufficient to calculate $(O-E)^2/E$ for the smallest value of O. If, as is the case here, the result for the smallest value of O (11) is (61-40.16 = 20.84) >3.84 (tabulated value for chi-squared at 1 degree of freedom and *$p = 0.05$)*, the chi-squared test is positive and the values A and B are associated.

The normal distribution

The standard deviation is especially applicable when the underlying distribution is close to normal (Gaussian), i.e. symmetrically bell-shaped. This is often assumed to be the case for many biological characteristics, among them height, weight and blood pressure. The normal distribution has some useful characteristics and many statistical tests can be used if the observations follow a normal distribution. Approximately, two-thirds of the values under a normal distribution curve fall within +/- one standard deviation of the mean, and approximately 95% fall within +/- two standard deviations of the mean (Figure 5-1). However, many biological distributions in parasitology and epidemiology do not follow a Gaussian (normal) curve.

Figure 5a: The normal distribution curve.



Exercises: Calculation of the mean, standard deviation and range

Exercise 1:

The duration of hospitalization for 24 children admitted last month to hospital X with pneumonia was as follows:

Child 1	06 days	Child 13	10 days
Child 2	07 days	Child 14	18 days
Child 3	10 days	Child 15	14 days
Child 4	08 days	Child 16	12 days
Child 5	11 days	Child 17	11 days
Child 6	08 days	Child 18	10 days
Child 7	04 days	Child 19	10 days
Child 8	17 days	Child 20	15 days
Child 9	15 days	Child 21	05 days
Child 10	14 days	Child 22	12 days
Child 11	08 days	Child 23	06 days
Child 12	11 days	Child 24	11 days

a. What is the range of values?

b. What is the mean duration of hospitalization?

c. What is the standard deviation?

d. Is range or standard deviation a better measure of the distribution in this case? Why?

Exercise 2:

The following data on the pulse rate were taken on admission from 10 male patients hospitalized last month:

83 beats/minute	59 beats/minute
72 beats/minute	72 beats/minute
77 beats/minute	58 beats/minute
62 beats/minute	65 beats/minute
60 beats/minute	77 beats/minute

a. What is the range of pulse rate values?
b. What is the mean?
c. What is the standard deviation?

Exercise 3 (chi-squared)

A sample of 200 shows the following:

Of 94 people with a positive blood slide, 34 regularly use insecticide-treated bednets. Of 106 people with a negative blood slide, 80 regularly use insecticide-treated bednets,

Tabulate the information and calculate the value of chi-squared. Is there a statistical association between the use of treated bednets and a negative blood slides ($p = 0.05$)?

Threshold values for chi-squared, $p = 0.05$ and $p = 0.01$

Degrees of freedom	$P = 0.05$	$P = 0.01$
1	3.84	6.64
2	5.99	9.21
3	7.82	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.01	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.03	26.22
13	22.36	27.69
14	23.69	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
8	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89

For instance, if the number of degrees of freedom in a distribution is 4 (as in the case of a 3 by 3 table), the threshold value for chi-squared at p = 0.05 is 9.49.

NOTES

Learning Unit 6

Principles of surveillance

Learning objectives

By the end of this Unit, you will be able to:

- Define the term *surveillance*
- Describe its uses in epidemiology and public health
- Describe the concept of a surveillance arc
- Describe the concept of feedback and why it is important
- Identify limitations of surveillance in drawing conclusions about health problems
- List the criteria for evaluating the usefulness of a surveillance system
- Apply these criteria to a country example.

Surveillance may be defined as the continuous collection, collation, analysis and interpretation of data on a systematic and ongoing basis, together with the feedback and dissemination of information to those who need it for action.

Information for ACTION

Uses of surveillance

Surveillance may be used at local, regional, national and international levels and may be used for purposes mentioned below (see also the surveillance arc on following page).

Providing baseline data

If appropriately collected, surveillance data can be used to determine baseline information on malaria deaths, and also on risk factors such as rainfall.

Goal setting

Surveillance can be used to gather information on prevalence and trends for the design of health interventions.

Assessing whether targets have been reached

If appropriately collected, surveillance data can be used to assess progress in reducing malaria deaths.

Example: *Assume a national goal has been set to "Reduce malaria deaths by 20% in the next five years". This implies that:*

- *there is a baseline figure on malaria deaths*
- *there is a way of monitoring trends in deaths over time.*

Detecting epidemics

If done in a timely fashion, surveillance will permit the detection of adverse trends before it becomes too late to alter their course.

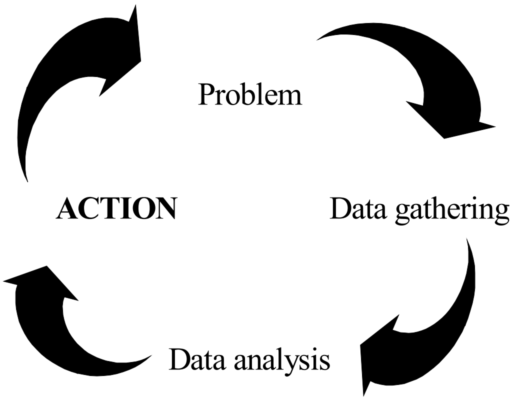
Identifying at-risk groups

Surveillance may be used to gather simple information on geographic subunits or age, sex or ethnic categories and identify areas or populations at increased risk of developing a disease or condition.

Some important concepts in surveillance

The surveillance arc

Figure 6a: The surveillance arc



Surveillance is meant to be **information** for public health **ACTION.**

Problems often occur with data analysis and action: data are gathered but not analysed or not summarized in useful form, or data are not disseminated to those who can make use of it, or the decision-makers are not aware of how the data can be useful to them. If data are to be used for decision-making, they must be made available and understandable to public health decision-makers in a timely fashion so that there can be feedback.

Timeliness

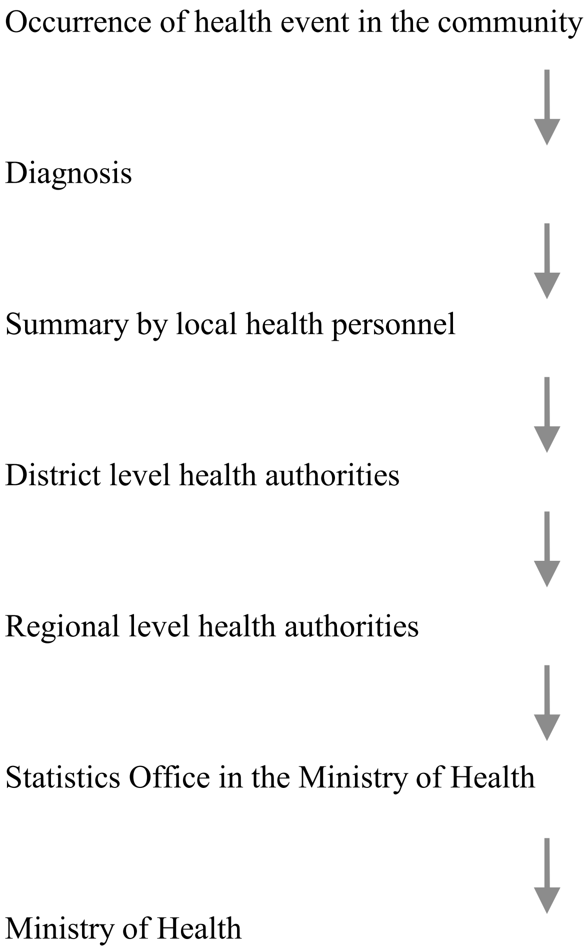
For data to be useful, they should be timely. Every effort should be made to send the data to the next highest level on a regular basis. Data should be examined and processed without delay at **each** level in the chain.

Feedback

For a system to operate well, those collecting the data need to receive feedback. Many countries have surveillance systems in place for a variety of diseases. The following is a schematic diagram of flow of information within many such surveillance systems:

Figure 6b: “Traditional” flow of information

What is missing in these situations is **feedback** to the previous levels within the data collection system, as shown in the following schema.



Occurrence of health event in the community

Diagnosis

Summary by local health personnel

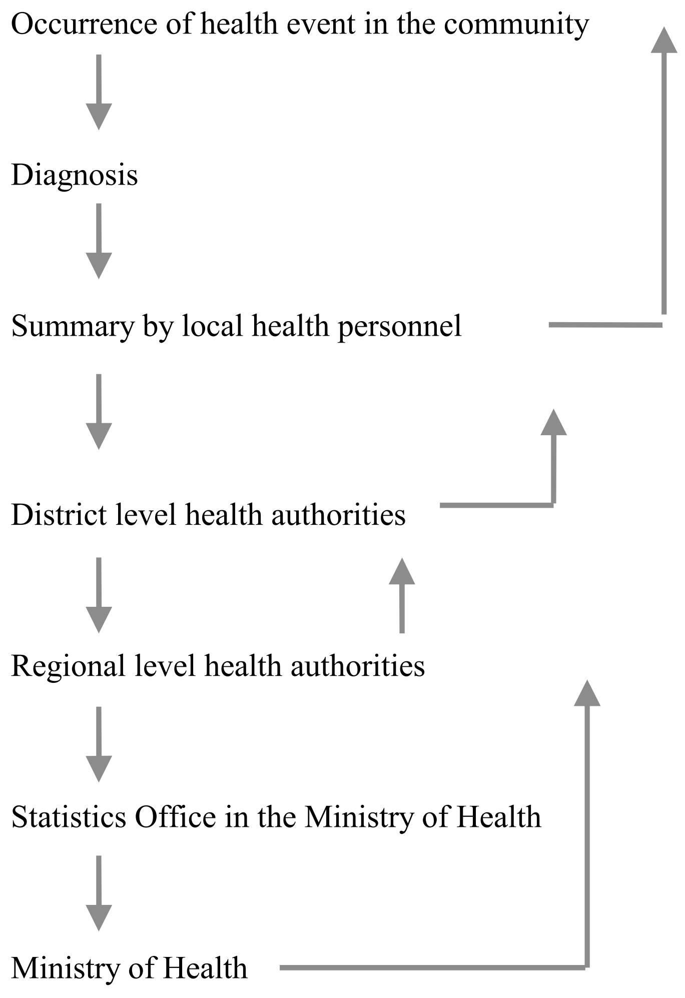
District level health authorities

Regional level health authorities

Statistics Office in the Ministry of Health

Ministry of Health

Figure 6c: Flow of information with feedback


Occurrence of health event in the community
Diagnosis
Summary by local health personnel
District level health authorities
Regional level health authorities
Statistics Office in the Ministry of Health
Ministry of Health

Feedback is essential to help those actually collecting the data realize that their data are being used and their efforts are worthwhile and appreciated. It can provide local health care providers with data that can be compared with data from other similar geographic subunits. Feedback is also useful for local or community planning purposes in countries where decision-making is decentralized. Because delays in getting the data analysed at the central level may be substantial, data must be summarized, examined, and acted upon at the first possible level rather than sending on to a more central level for analysis and processing.

The problem of underreporting

Surveillance data rarely include all the cases of a disease within a geographic area. Additionally, those cases covered by a surveillance system may not be altogether representative since they are based on people who voluntarily seek health care. The fraction of the total number of cases covered by the surveillance system and/or the representativeness of these cases may depend on:

- Availability of health services
- Severity of the illness
- Cultural perceptions concerning who should treat a disease (i.e. "modern" medicine versus traditional healers)
- Ability of health personnel to correctly diagnose the problem
- The diligence of health providers in reporting cases to the system.

Over-reporting may also occur (log-, double counting of cases). The interpretation of surveillance data therefore requires caution, and, generally speaking, surveillance data alone cannot be used to make broad conclusions about incidence or risk factors for a disease.

However, provided that biases remain roughly the same over time, it may be possible to use surveillance data in order to monitor disease trends. What is important is the slope of the line (the trend) rather than absolute numbers.

The problem of limited information

Because surveillance relies on routine data collection, only a limited amount of information can be collected. If you need more detail, you must carry out a survey and/or descriptive study (e.g. a chart review). This is often expensive and time-consuming, but much can be done with few data.

Example: *a system that collects data on cases and deaths by reporting site and age category will permit a variety of observations:*

- *Trends in cases*
- *Trends in death*
- *Trends in cases and deaths by age*
- *Seasonal variation of cases and deaths*
- *Geographic distribution of cases and deaths*
- *Case fatality rates.*

Numerator data may be useful for determining the distribution of health resources. However, to compare age groups, regions, etc. and to take into account the fact that populations generally grow over time, it will also be necessary to calculate rates. The rates calculated will usually represent an underestimate of the true rates because of underreporting within the surveillance system. Examples of items that can be calculated using data from a surveillance system include:

- Prevalence rates
- Trends in incidence rates (provided surveillance is ongoing)
- Trends in death rates
- Trends in case and death rates by age
- Geographic differences in case rates and death rates
- Other risk factors.

Evaluating a surveillance system [8]

A well-functioning surveillance system can be a useful public health tool for monitoring disease trends, determining areas and groups at risk, and monitoring the effectiveness of public health efforts. Although most countries have a disease surveillance system, many of these systems do not function optimally. The following paragraphs present a brief overview of four steps towards evaluating a surveillance system in order to improve the functioning of the system.

1. A first stage is to **describe the system** and outline how the data move through the system. To begin with, describe the objectives of the surveillance system (range of events under surveillance, population covered, intended use of surveillance data), and identify who should use the data and for what purposes. Next, set out the case definition (for surveillance purposes) for those conditions under surveillance. A flow chart of the system will show the progression of data through the system and indicate the levels at which analysis and feedback take place. Consider the following points:
    - What are the events under surveillance?
    - What is the population under surveillance?
    - What information is collected on each case?
    - Who determines whether a patient meets the surveillance case definition?
    - Who fills out the surveillance forms?
    - What type of feedback does each level undertake?
    - Who summarizes the data before they are sent to a more central level?
    - How often are the data sent to a more central level?
    - How is the information transferred?
    - How are the data analysed?
    - Who analyses the data, and at what intervals is it analysed?

[8] For a detailed discussion of this, see *Protocol for the evaluation of epidemiological Surveillance Systems*, WHO/EMC/DIS/97.2 (obtainable from CDS at WHO, Avenue Appia, Geneva, Switzerland).

- What is contained in the reports?
- How often are reports distributed?
- To whom are the reports distributed?

2. The second stage is to determine the usefulness of the system. Specifically, the following should be considered:
   - What actions are taken as a result of data from the surveillance system?
   - Who has used the data to make decisions and take action?
   - What actions have been taken on the basis of surveillance information (compare with anticipated uses of the data if possible)?

   If the usefulness of the system is low, it is worthwhile to find out why the information is not used and what may be done to make better use of it.

3. In addition to evaluating a surveillance system with respect to its functioning and usefulness, it is also helpful to look at the extent to which the surveillance system has the following attributes:
   - *Simplicity:* are the system and the flow of data uncomplicated and easy to carry out?
   - *Flexibility:* does the system have the ability to change to collect new or different information if the need arises?
   - *Acceptability:* is the system well accepted by the people who collect and use the data?
   - *Sensitivity:* does the system pick up a high percentage or at least a consistent percentage of the cases?
   - *Positive predictive value:* of the persons identified by the surveillance system as having the disease, how many actually have the disease?
   - *Representativeness:* are the findings from the surveillance system representative of the population covered by the system?
   - *Timeliness:* does the system work rapidly enough to provide useful information for decision-making?
   - *Uniqueness:* does the system unnecessarily duplicate other data collection efforts?

4. A final step in data gathering for the evaluation of a surveillance system is to assess the nature and amount of resources, both human and financial, required to operate the system.

Once all of the above factors have been examined, draw conclusions and make recommendations on how to improve the system. Specifically, the conclusions should discuss whether the system is meeting its objectives, whether it should be modified, or, in some cases, whether it should be discontinued. Recommendations should be of a practical nature and take into account limitations in funding and in human resources. Finally, once changes have been implemented, periodic re-evaluation of the system is useful in order to make further improvements. When changes will have been undertaken, it will be necessary to periodically evaluate the system towards other future improvements.

Suggestions to develop and maintain useful surveillance systems

- Decide why you need the data, who will use the results, and whether you have an **intervention** to offer.
- Decide what critical information you need and keep it **simple**.
- If universal routine surveillance is not feasible, try a good **sentinel** surveillance system where emphasis is on collecting high-quality data from a limited number of sites representative of the population.
- Process the data **without delay**.
- Feed results back to those who **collect** it as well as those who use it.
- Be **cautious** in interpreting data; keep in mind representativity (or lack thereof).

Exercises: Evaluation of surveillance

Exercise 1

While reading this exercise, use the following points as a checklist to identify problems in:

- Data collection
- Data analysis
- Use of information
- Relevance
- Feedback.

You have been assigned to the Office of Statistics within the Ministry of Health for a 3-month period. Your assignment is to make an evaluation of the existing surveillance system – the government is considering making the system more useful for decision-makers.

You decide to follow the data along the chain of reporting, beginning at clinic level, through the District Medical Officer (DMO) and Regional Medical Officer to the Office of Statistics and the Minister. You decide to examine the chain of reporting in Rado, since you used to work in the district and know many of the health officers there.

You meet with the Chief of the Statistics Office in the Ministry of Health, who says: "The system was established by the British, and we still use the same form we were using when I started working here 20 years ago. In the mid-1980s, Dr Obe from the University convinced the Minister to add 6 diseases to the list. We now collect data on 43 different diseases by sex for 3 different age groups".

Asked what problems he sees in the system, he replies, "Nobody seems to care much. It takes at least 9 months to get the reports from the regions. By the time we get all the data and put our reports together, they are out of date. I sometimes wonder if anyone ever makes use of them".

He tells you he will be happy to help you and gives you a copy of the most recent annual report, which was sent out last month to all district and regional medical officers. This contains data for 1993, even though we are now in 1998. It has over 200 pages, with tables of each disease by reporting site, sex, age, region and district; without graphs or text.

In Rado you visit a hospital, a clinic and two smaller rural dispensaries. All the health personnel you interview tell you that they learned how to fill out the forms as part of their initial training, but have no official guidelines on how to fill them out, and their supervisor rarely checks their work. You do an audit of the past month and find that for one of the dispensaries the figures in the report differ markedly from those in the register.

You ask the health workers what happens to the data after they are sent to the district level. All tell you they don't know; that is the last time they ever see the data. One of them tells you, "I think that the forms just sit on the DMO's desk and after a while he throws them away. This surveillance system is just useless paperwork that keeps me from more important work like seeing my patients".

You then visit the DMO. He tells you that filling out forms is a big part of his job. He fills out 8 forms a month (surveillance report, hospital reporting forms, essential drugs list, number of vaccines administered, sanitation worker report, supplementary food distribution form, hospital budget form, clinic budget form). Every month he plans on spending a little more time on surveillance but never gets around to it. He is aware that some of the clinics do not report routinely and that the numbers sometimes look suspicious, but he does not have time at the moment to fix the problem.

When asked how he uses the surveillance data, he admits he does not use them for anything. He shows you the 1993 report he just received from the Ministry and says, "How can you can make any sense out of this jumble of numbers?" Asked if he has ever used surveillance to detect an epidemic in the district, he says if there is a big outbreak he usually hears about it, although sometimes it is too late to do anything. He discovered a meningococcal meningitis epidemic last March when 11 cases showed up in the district hospital on the same day. When he later reviewed the surveillance records, he found that in January and February the number of cases had been 3 times higher than they had been in January and February of the previous year.

The Regional Medical Officer repeats many of the concerns expressed by the DMO. She states that the data might be useful for monitoring trends in priority diseases. Every time she looks at the annual report, however, there seem to be more and more cases of everything, even those diseases for which she knows there are good control programmes. Even if she could figure out what was going on, it is difficult to make programme decisions using information that is four years old or more.

Finally, the assistant to the Minister of Health tells you that he remembers from his primary health courses that surveillance is a good thing, However, he and the Minister do not really make use of surveillance data because they think that these data are old and rather unreliable. The information is rarely used in health planning for the country; they tend to rely more on disease prevalence data collected as part of a national health survey in

1995. He hopes that you will figure out a way to get more timely information, especially since this could be useful in following the progress of the new primary health care programme. As part of the bilateral funding agreement for this project, a modest amount of money will be set aside for improving disease surveillance.

Exercise 2

For three of the problems that you have identified, suggest a solution.

Learning Unit 7

Health facility-based epidemiological studies

Learning objectives

By the end of this Unit, you will be able to:

- Define the term *facility-based study*
- List the types of information commonly collected in descriptive studies
- Discuss the use of these studies and their limitations in drawing conclusions concerning the general population
- Describe the steps in conducting such as study
- Using sample data, draw up an analytic plan, analyse and interpret the results of a health-facility-based study.

For planning and other purposes, it is often useful to summarize data about the characteristics of persons coming to a health facility or people seen in that facility with a given diagnosis. Such studies fall in the general category of descriptive studies (see Learning Unit 1). They may be done either retrospectively using existing registers or charts, or prospectively using data collection forms that can be filled out when a patient with the disease under study comes to a health facility.

Most facility-based studies concentrate on describing the basic characteristics of cases according to **person, place,** and **time**. **Person** refers to characteristics such as age, sex, race, marital status, or occupation, but may also refer to behavioural characteristics of the individual, such as smoking, or whether a person has been vaccinated or has taken malaria prophylaxis. It can also refer to the patient's clinical or laboratory findings such as their white blood count or neurological status on admission, to the treatment they were given, or to their clinical outcome.

Place refers to geographic distribution and may include factors such as what district the person lives in, whether they live in a rural or urban area, etc.

Time refers, for instance, to the time at which the person contracted the disease or came to the hospital.

Occasionally, a study may be "nested" within a hospital-based descriptive study to study risk factors for poor outcome among people who have a given disease. For example, in addition to describing the characteristics of persons admitted with cerebral malaria, the health officer may wish to examine risk factors for death from cerebral malaria. In essence, this is a case-control study where the "cases" are those who have cerebral malaria and who die and the "controls" are those who have cerebral malaria but do not die. Alternatively, the health officer may wish to find out if the neurological status on admission predicts subsequent death from cerebral malaria. In essence, this is a cohort study where the "exposed" are those whose reactions are blunted or who are comatose and the "unexposed" are those who are awake and alert at the time of admission.

Record review

Most commonly, health facility-based studies involve the use of existing hospital or clinic records. With hospital records, for example, it is possible to examine the characteristics of children admitted with measles, including their age, sex, place of residence, complications developed, duration of hospitalization, and outcome.

The steps in a record-based study are usually as follows:

1. Determine the purpose of the study and the best source of data for the study. If a retrospective study is to be undertaken, examine a sample of records to determine the type of information available and its completeness.

2. Develop an analytic plan that includes all of the table shells that will be used in the analysis. This avoids the problem of collecting unnecessary information or of conducting analyses that are not pertinent to the purpose of the study.

3. Decide on the information required for each case and on what time period the study will cover. If there is a seasonal incidence to the problem, it is best to take cases from one or more years. If the number of potential cases is large, it is better to sample (i.e. take every fifth or tenth case) throughout the period than to select cases from only a few months.

   If the number of records is large, you will probably not need to review all of them. Use the formula for sample size calculations (see Learning Unit 8) to determine how many records you will need to review in order to achieve the level of precision you require. You may wish to sample different subgroups if you need accurate estimates for these groups.

4. Construct tally sheets with space to record information required. If feasible, it is best to try to enter the cases in a line listing format, with all the information relating to each case listed on a separate line, rather than recording each case on a separate form. This will facilitate data entry into the computer and the following

analysis. The table on neonatal tetanus (Exercises 1 to 4 in this Unit) is an illustration of such a line listing.

5. Identify registers or patients records that are needed from the health centre or hospital archives. It is important to find the records for *all* the cases chosen for inclusion in the study. Bias can occur if cases whose records are not found differ in some way from those whose records can be found, for instance if the records of the most severe cases cannot be found because they were transferred to a regional referral hospital. In this situation, reviewing the records that can be found would give a biased or distorted picture of the cases and their outcomes.

6. Review records and record the required information on the tally sheets.

7. Tabulate the data by hand or enter in the computer for tabulation into the shell tables previously developed.

8. Spend enough time looking at your results and draw your conclusions.

9. Prepare a written report as soon as possible, summarizing your findings.

Active data collection

Descriptive studies do not need to be passive; they can involve the deliberate collection of data on cases or on health care recipients. For example, say you need to know more about the characteristics of a given disease in order to allocate your resources more effectively, and your surveillance system or hospital records do not provide sufficient details on the cases. You may wish to develop a data collection form and collect more detailed information on all new cases of this disease admitted within a specified time period. Alternatively, you may wish to know about the characteristics of clinic attenders and how they perceive the services you are offering. In this case, you might develop a questionnaire to be filled in for all persons coming to your clinics during a specified period of time or for a sample of such persons.

Limitations of health facility-based studies

Hospital-based or clinic-based studies are useful to determine trends in morbidity and mortality, to establish the percentage of health consultations or hospitalizations or deaths from a specific disease, or to determine the characteristics of those attending with a given diagnosis.

Certain limitations must be kept in mind when performing health facility-based studies. Remember that this information is usually not suitable for calculations of actual incidence, prevalence or mortality rates, or to draw broad conclusions from the geographic unit under study, since these populations may not be representative of all those who live in the geographic unit and develop the disease. For example, it is difficult to draw conclusions from hospital data about the characteristics of children with diarrhoea since relatively few of these children will be hospitalized and since those who are hospitalized may not be representative of the entire population. If the purpose of a study is to obtain population-based rates of morbidity or mortality or to look at the characteristics of all those with a given health problem, it will usually be necessary to perform a survey or to undertake active case finding outside health care facilities.

Although "nested" case-control or cohort studies can be undertaken to look at risk factors for outcomes among those hospitalized with a given diagnosis, **you generally cannot use hospital-based studies to determine who is at risk of developing a disease or condition**. In order to determine risk factors, it is usually necessary to perform an analytic study in which the prevalence of the risk factor is compared between those who have the disease and those who do not have the disease. For example, a descriptive study may show that 80% of those with a disease belong to a given ethnic group and the investigators may erroneously conclude that belonging to this group is a risk factor for disease. However, this group may constitute 80% of the population coming to the hospital, and therefore this result may not correspond to an increased risk for the disease.

Summary

In summary, health facility-based studies can be used to:

- Examine trends in morbidity and mortality, and calculate case:fatality ratios.
- Describe the characteristics of patients coming for care to the health facility.
- Describe the characteristics of patients seen at the health facility with a given diagnosis or diagnoses.
- Examine risk factors for poor outcomes among **hospitalized** patients.

Health facility-based studies **cannot** be used to:

- Calculate incidence, prevalence, or mortality rates of a disease within the population.
- Draw broad conclusions about the characteristics of people with a given diagnosis or diagnoses.

Exercise: Health facility-based studies

The District Health Officer has reviewed the annual mortality statistics for the district hospital and finds that neonatal tetanus is one of the leading causes of mortality among the under 5 age group. He wants to know:

- Know what type of patients are admitted with this diagnosis and use this information to identify potential methods of prevention
- Look at risk factors for death among patients hospitalized with neonatal tetanus.

He therefore undertakes a chart review of the 22 cases of neonatal tetanus admitted to the hospital in 1994. A preliminary review of records shows that the data consistently available on each patient are:

- date of admission
- age
- sex
- presenting symptoms
- date of discharge
- outcome
- vaccination status of mother.

1. Using the available data, develop a plan of analysis.

2. Analyse the data according to your analytic plan (working in pairs).

3. Summarize your findings in words.

4. Can you draw definite conclusions about risk factors for contracting neonatal tetanus based on your findings? Why or why not?

NEONATAL TETANUS, ABABO HOSPITAL, 1994

	Name of Child	Chart No.	Date of Admission	Sex	Age at Admission	Delivery Site	Presenting symptoms	Date of Discharge	Outcome	Mother Immunized
1	JC	1004	2.1.89	F	7 days	Home	not sucking	14.1.89	died	no
2	LM	1213	17.1.89	M	3 days	Home	convulsions	18.1.89	died	no
3	TS	1311	31.1.89	M	13 days	Home	not sucking	28.2.89	lived	yes, one dose
4	AB	1446	5.2.89	M	4 days	Home	not sucking	14.2.89	died	no
5	NJ	1578	22.2.89	F	6 days	Clinic	not sucking	27.2.89	died	no
6	BB	1665	1.3.89	M	8 days	home	convulsions	3.3.89	died	no
7	MN	1756	16.3.89	M	15 days	clinic	not sucking	23.4.89	lived	yes, one dose
8	DR	1830	4.4.89	M	13 days	home	not sucking	28.4.89	lived	no
9	MP	1988	29.4.89	M	8 days	clinic	not sucking	7.5.89	died	yes, one dose
10	FD	2020	10.5.89	F	2 days	home	not sucking	13.5.89	died	no
11	DP	2143	27.5.89	F	6 days	home	convulsions	27.5.89	died	no
12	LD	2176	18.6.89	M	7 days	home	convulsions	23.6.89	died	no
13	JM	2238	23.6.89	M	3 days	home	not sucking	4.7.89	died	no
14	AN	2339	7.8.89	M	12 days	clinic	not sucking	21.8.89	lived	yes, one dose
15	JW	2648	2.8.89	M	9 days	home	convulsions	2.8.89	died	no
16	NM	2756	25.8.89	M	5 days	home	not sucking	9.9.89	died	no
17	AA	2909	3.9.89	M	6 days	home	not sucking	19.9.89	died	no
18	GB	2998	4.11.89	F	7 days	home	not sucking	29.11.89	died	yes, one dose
19	WP	3054	13.12.89	M	5 days	home	not sucking	25.12.89	died	no
20	EB	3112	13.12.89	M	12 days	home	not sucking	31.12.89	lived	yes, two doses
21	CL	3245	212.89	F	4 days	Home	convulsions	5.1.90	died	no
22	RY	3317	25.12.89	M	7 days	Home	not sucking	9.1.90	died	no

Learning Unit 8

Surveys and data management

Learning objectives

By the end of this Unit, you will be able to:

- Define the term *survey*
- List the types of information commonly collected in surveys
- Discuss the use of surveys and their limitations
- Define *complete survey*, *simple random sampling*, *systematic sampling*, and *cluster sampling*, and describe their advantages and disadvantages
- Describe the steps in conducting a survey and the importance of developing an analytic plan as one of the first stages therein
- Develop an analytic plan for a survey
- Describe the steps in questionnaire development and testing and the ordering of questions within a questionnaire
- Describe the advantages and disadvantages of open-ended versus closed questions in surveys
- Develop a questionnaire
- Select a sample from a population list
- Examine and interpret data from a field survey

An overview

A survey may be defined as the collection of information from all individuals or a sample of individuals chosen to be representative of the population from which they are drawn.

Types of information collected by a survey

- Morbidity prevalence
- Morbidity incidence (generally collected by inquiring about events that have occurred sometime in the past)
- Mortality (also generally collected by inquiring about deaths that have occurred in the past)
- Detailed risk factor or behavioural information
- Knowledge, attitudes, and practices
- Physical signs (paralysis, splenomegaly, malnutrition)
- Serological or laboratory tests.

Characteristics of surveys

- **Representative** if sample chosen correctly.
- Single point in time - **snapshot**.
- Provide more in depth information than surveillance or chart reviews.
- Usually performed by a limited number of personnel specially trained to perform surveys.
- Can sometimes be expensive, time-consuming to perform.
- Cannot be used to monitor change unless repeated at a later time; may therefore be more suitable for situational analysis than for ongoing monitoring of a current problem or a programme.

When to undertake a survey

- When accurate population-based data are needed to determine the magnitude of a problem
- When more detailed or recent information is needed than is available from record review or surveillance (demography, examination, laboratory)
- When information is needed on health problems that may not routinely be seen by health providers
- When information is needed on health behaviours or health knowledge and attitudes, and this information is not routinely available through existing mechanisms.

Major steps in conducting a study

GENERAL PREPARATION

A total of 16 steps have been listed in the box and will be reviewed in detail, with relevant exercises, until the end of this learning unit.

PREPARATION	
Step 1	**Determine the objectives of your study**
Step 2	**Determine exposure and outcome variables**
Step 3	**Develop preliminary "shell" tables**
QUESTIONNAIRE	
Step 4	**Design a questionnaire**
Step 5	**Field test questionnaire in the population in which it is to be used**
SAMPLING	
Step 6	**Determine the study subjects, the method used and the sample size required**
Step 7	**Establish a sampling plan**
LOGISTICS	
Step 8	**Determine personnel needs**
Step 9	**Develop instruction manuals**
Step 10	**Select and train the personnel to be used to collect the data.**
Step 11	**Develop a checklist of logistics**
FIELDWORK	
Step 12	**Collect the data**
Step 13	**Edit data**
ANALYSIS	
Step 14	**Analyse the data**
Step 15	**Interpret your data**
REPORT WRITING	
Step 16	**Write up results NOW and disseminate them to the appropriate people.**

A. PREPARATION (1-3)

Step 1
Determine the objectives of your study

- What question(s) are you trying to answer?
- Who will be using your findings?
- How will these findings be used?

Step 2
Determine the exposure and outcome variables you will study and decide how you will define them

- Sources: literature, experts, focus groups, preliminary interviews.
- Justify the inclusion of each variable.
- Avoid the temptation to include variables that "might be interesting".
- Realize you may need more than one study.
- Decide how variables are to be classified.

Step 3
Develop preliminary "shell" tables (dummy tables)

- Begin with simple descriptive characteristics.
- Develop shells for two-way tables.
- Develop shells for other tables.

Exercises: Preparation

In Mariabo, central Africa, a review of childhood mortality data shows that measles, diarrhoea and malaria are the leading causes of death. The rates of child death from measles and diarrhoea are declining, but the rate of death from childhood malaria is increasing. As Director of the national malaria programme, you are asked by the Minister to determine what is happening and how these deaths might be prevented.

In the past 5 years, you have launched a campaign to make chloroquine widely available in the street markets. You created a series of posters for Maternal and Child Health clinics instructing mothers to give chloroquine when their child has fever and stating how many tablets to give each day and for how many days. Physicians at the hospitals who take care of children with malaria estimate that less than 10% of mothers are treating their children as recommended, and that if the mothers do treat their children, they are not using the right dosages.

You are not sure whether the problem is due to gaps in knowledge or to gaps in practice. You decide that a survey of maternal knowledge and practices on the use of chloroquine in children with fever would be useful to better direct your efforts towards the reduction of childhood mortality from malaria.

Exercise 1:

Develop a series of objectives for the survey (knowledge and practice).

For one of these objectives, list

- at least 3 maternal demographic variables
- at least 2 knowledge variables that will be useful in planning an intervention programme.

Exercise 2:

Develop a series of shell tables using these variables.

B. QUESTIONNAIRE (4-5)

Step 4

Design a questionnaire that will allow you to "fill in the blanks" in the shell tables(the analysis should drive the questionnaire rather than vice-versa).

Step 5

Field test the questionnaire in the population in which it is to be used and determine whether there are operational problems. Revise the questionnaire/ methods as needed. If necessary, develop other survey forms (record-keeping forms for interviewers to keep track of sites visited, etc.). When the questionnaire is finalized, develop a data entry programme if computers are to be used.

Questionnaire design and construction

The questionnaire must be written after the analytic plan is developed and must be designed to collect the information required for analysis. Review pertinent information from literature and experts to ensure that you are collecting appropriate variables.

- A decision needs to be made early as to whether
  the respondent will be interviewed or
  he or she will actually fill out the questionnaire.

This will have a major impact on questionnaire language, design, and layout.

Whether questionnaires are filled out by the respondent or an interviewer, they must be appealing to the eye and easy to complete and code. Questionnaire appearance affects response rate as well as the ease of data summary and analysis.

Research and use existing questions from other well-conducted surveys. This allows you to take advantage of others' experience and also promotes comparability of data collected by different groups.

There are several basic types of questions

Open-ended

Example: What do you think of the care you receive at the clinic?

Multiple choice

Example: Which of the following describes your marital status?
a) Married
b) Single
c) Widowed, separated or divorced

Fill in the blank
Example: My choice for prime minister would be.........

Rating scales
Example: Are you satisfied with the care you receive at the clinic?

0	***1***	***2***	***3***	***4***	***5***
(Dissatisfied)					***(Very satisfied)***

The type of questions to be used will depend on the purpose of the question, the size of the study, and who will be filling out the questionnaire. In general, however, keep open-ended questions to a minimum unless doing a small or exploratory study. Although they may provide more detailed information, they will usually be hard to code and subsequently analyse.

Identifying information (e.g. name of town, hospital etc.) must be part of the pre-coded information on the form. Ask the respondent only for information you do not already have.

Word the questions as simply as possible, taking into account the education level of your respondents.

Use familiar and appropriate language. Avoid slang, abbreviations, double negatives, and emotionally laden words.

Example: When was the penultimate occasion you took your offspring to an ambulatory facility for anthropometric assessment? is a confusing (and pretentious) way of saying what can be stated much more clearly as: *When was the last time you took your child to the clinic to be weighed and measured?*

Example: Don't you think that people should not smoke? This contains a double negative. Reword the question as*: Do you think people should smoke?*

Ask only one question at a time.

Example: Do you smoke and drink? This must be divided into two questions:
Do you smoke?
Do you drink?

Avoid language that suggests an answer.

Example: In your view, are authorities doing enough against malaria? This may suggest to people that the interviewer wants a "no" answer. It is better to ask an open question such as: *Are you satisfied with what is being done against malaria?*

Avoid hypothetical questions.
Example: If you were to become pregnant, where would you seek pre natal care? There is no ideal alternative, but if a woman has had more than one pregnancy, you could ask: *For your last pregnancy, where did you seek prenatal care?*

Be sure the respondent knows whether you want a factual answer or an opinion.

Example: The person who is asked *"What do you know about malaria control programmes?"* may not know whether he or she is being tested on knowledge regarding which is the better option or whether you simply want to know what he or she thinks. It is better to ask: *Are you satisfied with what is being done against malaria?*

Options for answers

Provide sufficient categories of answers
Be sure categories are mutually exclusive
Example: With the following question, people may fall into more than one category:

Has your child ever had:

(a) diarrhoea
(b) malaria
(c) measles
(d) pneumonia

It is better to ask:

Has your child ever had:

(a) diarrhoea	*Yes*	*No*
(b) malaria	*Yes*	*No*
(c) measles	*Yes*	*No*
(d) pneumonia	*Yes*	*No*

Provide “unknown”, “other”, or “don't know” options where appropriate. Use "unknown" only if you do not want to force your respondent to choose one of the other categories.

Arrange responses vertically, and mark "fill-in-the blank" information.

Provide the unit of measurement for continuous variables.

Facilitating the data processing of questionnaires

Pre-code the response categories. If the form is being filled out by trained interviewers, have them write the code directly in the appropriate box.

Provide a space for each digit you expect the respondent or interviewer to code, e.g. _ _ (month) rather than _____ (month).

If using interviewers, place a column along the margin, in which the answers can be coded close to the question; this will greatly facilitate computerized data entry.

Use the same coding scheme throughout the questionnaire; i.e., code none or “absent” as 0, 00, or 000, code unknown as 9 or 99 or 999 for all questions; code other as 8 or 88 or 888, for all questions.

Questionnaire order

Arrange questions in logical sequence.

Group questions by topic; place a few sentences of transition between topics.

Begin with easy and non-threatening but necessary questions. Threatening topics are often culturally determined and may include sex, violence, religion, patriotism, race and minority groups, use of illicit drugs, politics, money. Place these questions towards the end of the questionnaire.

Use skip patterns where appropriate (e.g., “If your answer to this question is “no”, go to the following question. If it is “yes”, go directly to question 24.”).

When collecting residential, reproductive, job, or other histories, follow chronological order either forward or backward in time.

Include "Thank you" after the last question.

Questionnaire format

Have the title of the study on the front of the questionnaire in **bold type**. The data and name of the organization conducting the study should also be on the front for identification.

Arrange the questions on the page such that there is adequate space between them.

Number the questions and letter the possible answers. This will prevent questions from being inadvertently overlooked and to facilitate the use of skip patterns. Answers must be indented.

For open-ended questions, provide sufficient space for the answer.

Keep instructions clear, brief, and precise. Print them in **bold type** or ALL CAPITAL LETTERS or *italics* and/or place in a box. Print skip instructions close beside relevant questions.

Be consistent. Use either "circle number" or "check box" but not both in one questionnaire.

If the questionnaire is printed on both sides of a page, put "over" or "PTO" on both sides. If the questionnaire is longer than two pages, use a booklet format for ease of reading and turning pages, and to prevent loss of pages. Make sure individual identifiers occur on each page of a multi-page form.

Do not split a question between 2 pages, since the respondents/ interviewers may think the question is completed at the end of the first page.

When asking identical questions about multiple household members or events, use a chart with parallel columns and facing pages if necessary.

Pretesting the questionnaire

Ideally, the following steps must be followed before beginning to conduct a study:

Obtain peer evaluation of the questionnaire.

Revise and test the next draft on yourself, friends, relatives, and co-workers, and develop a revised draft for field testing.

Prepare simple interviewer instructions for the field test.

Field-test the questionnaire on 20-50 respondents similar to the population you will be interviewing in your study.

Obtain the comments of interviewers and respondents concerning the questionnaire.

Eliminate questions that do not discriminate between respondents or that do not appear to provide the type of information required.

Revise and field-test again.

Prepare final interviewer instructions.

During interviewer training and initial interviewing, be alert to possible new problems.

If it is impossible to carry out all these steps, the questionnaire must be pretested, at a minimum, on a group of subjects similar to those that will be included in your study.

Exercises: Questionnaire

Exercise 3:

With your group, write a series of questions designed to find out:

(a) whether the child had a fever in the past month

(b) whether the mother sought medical attention for this episode

(c) where she sought the treatment

(d) if she had not sought medical attention, whether she treated the fever herself with chloroquine

(e) where the mother obtained the chloroquine

(f) whether she had given an appropriate dose for the right amount of time

(g) how far the mother lives from the nearest health centre

C: SAMPLING (6-7)

Step 6

Determine who will be the study subjects, the method(s) you will use to gather the data and the sample size that will be required using this method, taking into account statistical considerations and limitations of cost and of personnel.

Step 7

Establish a sampling plan for data collection and work out the logistics. Establish the method to be used for collecting the information, including types of questionnaires, day-to-day activities, plans for specimens, data entry, etc. Software such as Epi-Info is well suited for data entry and analysis.

Selection of samples

The following represent the types of sampling most frequently used in health surveys:

Complete or **comprehensive survey** in which measurements are taken from each unit in the population. This may be used if the total number of individuals is small (e.g. nurses in a single hospital or all patients admitted with a given diagnosis over the course of a year). It avoids problems with sampling and eliminates the need to calculate confidence limits or standard errors, but you may waste precious time and resources if the survey includes more individuals than are actually needed for the required degree of precision.

Probability sample survey in which measurements are taken from representative units selected from the population. The most commonly used types in developing countries are systematic sampling and cluster sampling. Systematic sampling is often used in reviews of hospital records or in studies of health care workers, while cluster sampling is most frequently used in surveys of widely dispersed populations.

If it is important to have precise estimates of a value for different subgroups within a population, the sample size will need to take this into account (this is called **stratified sampling**). For example, you may wish to look at the five leading causes of admission in each of the five hospital services. If the obstetrics service is only half the size of the other services, you may wish to take a different sampling ratio for the obstetric service in order to ensure sufficient accuracy in your estimates.

Systematic sampling

This may be defined as a type of sampling in which all units in the population (health workers, children, women, households, etc. depending on what is being studied) are counted and numbered and every *n*th unit is included in the survey. The *n* is dependent on the number of units and the required sample size.

Steps in selecting a systematic sample

1. Determine the total number of units in the population. If the units are not already individually numbered, they will need to be numbered before the actual survey can begin.

2. Determine the sampling interval K:

$$K = \frac{\text{number of units}}{\text{desired sample size}}$$

3. If the sampling unit is the child under 5, you will need to first figure out how many houses must be visited in order to find one child under 5, usually through doing a pilot study of 30 or so families. Similarly, if the sampling unit is the pregnant woman, you would need to figure out how many households to visit to find a pregnant woman.

   Example: Assume you are doing a study involving children under 5. There are 1500 households in all, and you have a required sample size of 100 children. From a preliminary study you have done, there is one child every 2.5 households. Normally, if there were a child in every household, you would visit 100 households. But because not every household includes a child, you will need to visit 100 x 2.5 or 250 households to find the required 100 children. The sampling interval will therefore be 1500/250 or every 6th household.

4. Select a number between 1 and the sampling interval for your starting point.

5. Add the sampling interval to the chosen starting point to obtain the second sampling unit. Add the interval to the second unit and so on until the entire sample is selected.

This is a useful technique only if:

- The population can be easily enumerated.
- The population is not widely dispersed.

The technique is easy to apply and avoids bias due to the data collector, but cannot be applied if the individuals are seen in a non-random order or if they are coming to the sample of their own volition.

Cluster sampling

Cluster sampling is defined as a procedure in which groups of individuals rather than the individuals themselves are the units of sampling. In this technique, each group of individuals is considered as one sampling unit. A cluster may consist of individuals living in the same household, individuals living in a neighbourhood, individuals working in health centres or hospitals, school classes, or more commonly, in the same village or town. The likelihood of a cluster being chosen for inclusion in the sampling is proportionate to its size. Depending on the sampling units used, it is often then combined with a second method of sampling for selecting individuals within the cluster for inclusion in the study.

Steps in selecting a cluster sample:

1. Enumerate all population concentrations in the sampling universe.

2. Draw up a cumulative population list.

3. Determine the sampling interval by dividing the overall population by the number of clusters desired (the ideal number of clusters is about 30).

4. Pick a number between 1 and the sampling interval from a random number table to use as the starting point and sequentially add the sampling interval until the desired number of clusters are obtained.

5. Once clusters are obtained, go to the site and pick the required number of individuals. The number of individuals in each cluster is obtained by dividing the desired sample size by the number of clusters. Although cluster sampling does not require an actual list of all the individuals living in a population, it does require a reasonably accurate population list. Old census data may be used if the increase in population has been relatively uniform in all areas, but if there has been a massive migration, war, etc, these data may be unreliable. If there has been no good census, consider other population lists (tax rolls, population lists kept for political purposes, etc.) or do an ad-hoc census with the help of local leaders (since it may be hard for the leaders to come up with actual numbers, the census may be qualitative in this case rather than quantitative e.g. + to ++++ to describe the size of villages or towns).

6. Another method of sampling is **quota sampling**, in which houses are visited sequentially until the required number of individuals has been obtained. The method must be decided upon ahead of time and used consistently.

Determining sample size

The general concept of determining sample size is that a study must be large enough to provide reliable estimates, but no larger than needed, so that resources are not wasted.

- Establish the methods to be used for collecting the information (questionnaires, day-to-day methods, plans for specimens, etc.).

- Train the staff and field-test the questionnaire and other data collection methods.

- Draw survey maps, arrange logistics and supervision.

- Conduct survey.

The factors to be considered in determining sample size for studies involving estimations of incidence or prevalence include:

- Sampling universe (what is the population you are sampling from?). You must decide on the geographic area to be surveyed, as well as who in the population is to be included in the survey (e.g. children under 5, women of reproductive age, clinic attendees at a given clinic).

- Sample size (see relevant section below).

- Sampling method (see below).

- Personnel

- Resources

Statistical considerations

- The prevalence *p* of the characteristic in the population is expressed as a percentage (estimated from previous or pilot studies; if totally unknown, 50% gives the largest possible sample size).

- The maximal acceptable difference of estimate from true value *d* (example: if the true value is 40%, is it acceptable to say that the true value is somewhere between 30 and 50% (*d* = 10%) or do you want to be able to say that the true value is between 35% and 45% (*d* = 5%)?

- The acceptable margin of error (usually 95% or 90%) that the true value does not lie outside the range chosen *Z*; for 95% the value for *Z* is 1.96, while for 90%, it is 1.64.

- In practice, many population-based surveys are cluster surveys. In this case, the design effect (*DEFF*) needs to be taken into account. In malaria surveys this ranges from 4 to 6, so all estimates on sample size need to be multiplied by 4 to 6. If a cluster survey is performed and a design effect of 5 is assumed, the value of Z^2(p)(100-p) would have to be multiplied by 5.

Formula $n = Z^2(p)(100\text{-}p)xDEFF/d^2$

- How do you know what values to use for *p*, *d*, *Z* and *DEFF*?

- The value of *p* is estimated from other surveys or from educated guesses. If unsure, you may wish to calculate a sample size based on a range of possible estimates (e.g. between 25 and 50%).

- The value of *d* depends on how precise you feel you need to be. If you are planning to repeat a study at a later time and want to look at differences in prevalence, you will want a smaller value of *d* to minimize the risk of the range of values for the two prevalence figures overlapping. Since *d* is squared, however, even a small change in the desired precision will considerably increase the sample size required.

- *Z* is arbitrary, but is usually set so that you want to run a risk of less than 5% of cases that the true value (if you could sample the entire population) rests outside the sample prevalence plus or minus *2d*).

- *DEFF* is estimated from other similar studies. In general, for nutrition, family planning, KAP and immunization surveys, it can be assumed to be 2. Diarrhoeal diseases and malaria require higher design effects, since people within a village tend to be relatively homogeneous with respect to these two diseases.

Example: Suppose you wish to perform a systematic sample survey to estimate the prevalence of parasitaemia in children under 5 in a district. Based on previous surveys, the estimated prevalence (p) is 15%. You decide to tolerate a less than 5% chance (Z=1.96, in practice 2) that the true value is smaller than 10% or larger than 20% (d=5%). Because it is a systematic sample survey rather than a cluster survey, the design effect becomes 1. Applying the formula, the following result is obtained:

$$1.96^2(15)(85) \times 1/5^2 = 196$$

For studies comparing rates or averages between two groups, different formulas are needed. See statistics textbook, or a statistician. The same applies if you wish to take into account the cost of the survey in deciding upon the sample size.

Miscellaneous comments

Sample size also needs to take resources and cost into account. In such cases, seek the help of a statistician, remembering that all efforts must be made to obtain the required sample size. Sometimes the sample size can be increased by decreasing the amount of information obtained from each subject, so that the same workers can interview more subjects.

If the number of individuals in the sampling universe is very small, you might decide not to sample at all but to take everyone.

If the total population you are studying is less than 10 000, or when $n > 01.N$, a finite population correction must be taken into account. This is done using the following formula:

Corrected sample size = $n[1-(n/N)]$

where n is the sample size calculated using the previous formula and N is the size of the total population.

> *Example: Assume that the district in which you wish to do a systematic sample malaria prevalence survey has only 1000 children under 5. The corrected sample size is 196[1-(196/1000)] or 158. In this situation, only 158 children under five would need to be included in the sample.*

The prevalence figure used is that of the most important variable(s) you wish to learn about. Think carefully about what is important.

> *Example: In the above-mentioned malaria study, the team decided that in addition to monitoring the overall prevalence of parasitaemia, they also wished to have a reasonably reliable estimate of the percentage of parasitaemic children who had been symptomatic and had been taken to the health clinic. They estimated that among the 15% of children they estimated to be parasitaemic, about half would have been taken to the doctor. Thus, the prevalence that they wanted to look at was half of 15% or about 7%. They then decided that they needed to know with 95% certainty that the true value would be plus or minus 2% around the prevalence obtained, and the sample size increased to 625 $[1.96^2(7)(93) \times 1/2^2 = 625]$. Note that this will have the side result of giving a more precise estimate for the levels of parasitaemia, but at the expense of considerably enlarging the survey size.*

Instructions on use of random number table

- Determine the total number of digits contained in the sampling interval you have calculated (example, 15 – with 2 digits, 1000 – with 4 digits, 26 980 – with 5 digits).

- Arbitrarily choose the direction you will follow within the random number table (right, left, up, down).

- Close your eyes and touch a pencil point to a number on the random number table.

- Begin counting off the required number of digits from the first step above in the chosen direction until you find a number between 1 and the sampling interval. Note that for numbers such as 1005 or 21 300, you may need to count off many numbers until an appropriate one is found. If you come to the end of the row or column, loop back in the opposite direction in the next row or column.

Table of random numbers

18 10 49 89 75	57 96 23 76 80	93 00 28 92 31	44 33 49 42 80
50 89 75 71 55	27 63 29 98 47	38 94 60 09 62	61 42 86 50 58
11 15 50 84 49	34 67 34 36 82	53 90 49 23 88	06 89 27 08 16
70 25 51 01 81	16 19 30 09 68	02 21 05 62 33	45 95 87 67 47
62 86 38 01 20	04 82 62 77 31	49 63 64 70 99	39 66 55 18 11
95 19 70 36 92	85 05 39 25 78	84 34 14 28 76	20 20 17 79 94
85 61 50 19 61	87 14 59 61 75	53 44 19 12 00	65 02 00 70 99
83 55 66 76 74	68 47 68 66 86	49 47 63 51 43	87 42 58 36 04
90 51 34 31 18	74 55 41 42 81	70 15 36 55 16	10 88 62 68 72
99 56 78 99 98	77 87 25 77 60	34 13 82 02 11	32 31 43 48 10
27 24 80 09 77	14 13 96 19 16	22 48 88 26 25	42 67 93 74 00
34 63 66 89 97	29 99 91 27 17	14 56 41 05 32	90 14 45 30 61
28 98 45 23 35	60 68 32 66 37	43 44 27 92 07	91 64 22 32 72
06 96 34 21 67	08 12 58 74 35	91 64 68 15 01	36 52 07 00 39
19 62 94 14 54	83 15 22 30 16	92 99 79 27 67	13 22 25 43 19
44 36 96 82 39	55 96 96 89 04	43 89 96 59 17	10 84 24 12 44
76 96 59 93 98	79 41 35 91 77	66 88 50 31 77	06 24 08 19 51
31 61 97 08 88	35 43 85 84 51	94 85 55 05 33	86 42 20 51 41
42 95 12 75 72	33 23 70 66 71	76 89 28 45 92	12 21 41 92 53
95 42 30 03 62	83 35 78 07 35	67 85 83 57 36	96 97 62 67 06
48 55 12 87 21	41 86 33 99 44	83 14 01 42 54	59 31 64 10 04
46 18 81 87 56	81 03 74 48 49	28 37 85 93 69	84 92 33 52 70
66 47 43 88 02	61 25 59 10 35	09 65 92 36 93	47 04 89 17 03
61 91 88 50 00	19 31 08 80 39	14 03 80 46 41	78 82 03 69 52
85 74 04 57 53	44 43 44 61 57	29 24 36 38 79	49 25 39 73 02
89 09 53 94 07	92 21 54 01 70	31 91 39 51 03	94 83 98 31 15
54 87 27 50 35	73 27 60 10 55	13 21 24 10 55	84 78 88 46 83
49 13 89 98 96	21 02 44 94 30	50 70 71 02 16	35 31 13 14 45
97 37 11 88 77	45 16 03 17 01	00 67 28 09 39	28 39 11 36 82
99 70 37 54 02	40 71 13 59 37	84 38 47 11 31	48 92 28 96 37
65 67 36 23 39	07 20 59 36 85	47 17 51 32 75	07 74 63 68 01
53 69 94 34 45	46 09 52 84 40	82 80 75 72 79	43 97 07 96 15
54 08 33 44 54	42 81 46 46 42	01 44 13 13 97	35 11 85 48 41
95 54 39 60 78	27 35 07 35 53	93 29 83 01 86	52 11 41 68 50
88 79 66 20 03	48 81 94 46 07	91 39 12 45 51	68 94 53 77 83
68 82 57 41 23	57 52 47 09 83	11 27 88 40 16	22 64 86 22 18
55 73 62 41 71	45 35 51 28 64	82 46 10 85 71	21 57 92 10 58
17 50 60 03 20	35 64 36 90 97	29 78 17 83 29	08 99 20 47 79
11 64 11 75 35	76 49 67 96 84	11 75 73 34 90	97 74 85 B8 37
78 32 11 34 33	55 30 20 68 10	68 96 94 82 04	94 10 52 73 51

Exercises: Sampling

Exercise 4

Describe possible methods for sampling in your survey of maternal knowledge and practices.

Exercise 5

You have decided to conduct a cluster survey using 30 clusters.

How big a sample will you gather ?

Exercise 6

When you go to the Ministry of Planning to obtain the most recent census,
you find that this was last done in 1981.

Can you use a census this old for your sampling? What are the alternatives?

Exercise 7

Calculate the cumulative population for the first 5 towns in the list of villages, towns, and cities in the region (page 107)

From the list, select the sites to be included in the survey.

Exercise 8

Describe how you will select the individuals in each site and what you will do if there is more than one cluster selected from a single location.

Population list based on 1981 census, Mariabo. Villages/towns have been given numbers rather than names for purposes of this exercise. You must calculate the cumulative population for the first 5 sites.

SITE	POPULATION	CUMULATIVE POPULATION
1	2 140	……….
2	15 757	……….
3	4 148	…….…
4	1 732	……….
5	2 506	……….
6	2 171	28 454
7	29 098	57 552
8	1 092	58 644
9	4 973	63 617
10	1 884	65 501
11	957	66 458
12	4 907	71 365
13	3 009	74 374
14	14 871	89 245
15	59 895	149 140
16	7 587	156 727
17	2 371	159 098
18	557	159 655
19	2 909	162 564
20	28 771	191 335
21	2 588	193 923
22	1 651	195 574
23	2 839	198 413
24	9 916	208 329
25	5 653	213 982
	213 982	

D: LOGISTICS, FIELD WORK, ANALYSIS, REPORT (8-16)

Step 8

Determine the personnel needs (types of people and necessary person-hours) and develop appropriate descriptions of responsibilities for each level of personnel (supervisors, surveyors, drivers/guides, translators).

Step 9

Develop instruction manuals for survey personnel detailing how questionnaires are to be filled out and how the sample is to be selected. Determine how field supervision will be performed.

Step 10

The keys to training are information, examples, and practice. Select and train the personnel to be used to collect the data.

Step 11

Develop a checklist of materials needed for field work (forms, papers, pencils, clipboards, paperclips, sleeping bags, tents, etc.). Arrange per diems, payment schedules, daily rates of pay and of compensation, vehicles, drivers and other logistic elements.

Step 12

Collect the data, assuring their quality and completeness through supervisory visits and review of data forms.

Step 13

Edit your data to eliminate errors in collection, coding, transcription, or data entry.

- If data are entered in the field, build in edit checks.
- Otherwise this is done by checking data for each variable and looking for abnormal values, extreme ousters, and unexpected population distributions.
- Perform plausibility edits (children dead before they were born, etc.).

N.B. If errors are found:

- Go back to the source wherever possible.
- Avoid second-guessing.
- Be consistent.
- Correct an error as soon as it occurs.
- Document the correction.

Step 14

Perform the data analysis

- Calculate the response rates.
- Fill out the skeleton tables you developed, starting with the simple tables and proceeding with the more complex.
- Reduce the data if necessary by collapsing categories (but do not manipulate the data to improve your results!!!).
- Look long and hard at your data and think about what they mean (this may take hours…).
- Perform necessary measures of association and statistical tests keeping in mind your chosen study design (e.g. matching, design effect).

Step 15

Interpret the data

- Spend time thinking about the meaning of the results.
- Be cautious in interpreting: Non-significance does not necessarily mean "no association". Conversely, significant results may have no public health importance. Remember that statistical tests do not show a cause-effect relationship; they just rule out the role of chance in the findings.

Step 16

Write up results NOW and disseminate them to the appropriate people.

Do the analyses yourself whenever possible – it gives you a better feeling for the data, not to mention the thrill of discovery.

Exercise: Analysis

Exercise 9

Using the data below, summarize the findings and make recommendations for improving the current policy of encouraging early home treatment of fever with chloroquine.

Number of children under 5 included in the study = 860
Number experiencing fever in the last month = 503
Number of mothers seeking medical help for the fever = 207/503

Number of mothers not seeking medical care for treatment of fever = 296/503

Number of children whose mothers did not seek medical care but did initiate chloroquine treatment on their own = 39/296

Number of children who were neither seen by a health worker nor treated with chloroquine by mother = 503 – 207 – 39 = 157

Number of mothers seeking medical help for the fever by distance from the health centre

Distance	seeking help/total
1 hour	93/194
1-< 2 hours	54/135
□ 2 hours	27/65
unknown	33/109

Number of mothers seeking medical help for the fever as a function of the child's sex

Sex	seeking help/total
males	100/256
females	106/246
information missing	1

Number of mothers seeking medical help for the fever as a function of the child's age

Age	Seeking help/total
< 1 year	39/ 95
1 year	53/116
2 years	46/117
3 years	47/110
4 years	22/ 65

Among the mothers initiating treatment on their own (n = 39)

Source of medication	frequency/total
pharmacy	12/39
market	8/39
health agent (from previous illness)	7/39
relative or friend	5/39
other	7/39

Total dose given (available for 29 children)

11/29	received more than 30 mg/kg	(37%)
9/29)	received between 20 and 30 mg/kg	(31%)
9/29	received less than 20 mg/kg	(31%)

mean dose received:	32.7 mg/kg
median dose received:	26.7 mg/kg

(The recommended dose is 25 mg/kg, given in 4 doses at 0, 6, 24, and 48 hours)

NOTES

Learning Unit 9

Assessing the accuracy of a test or surveillance system

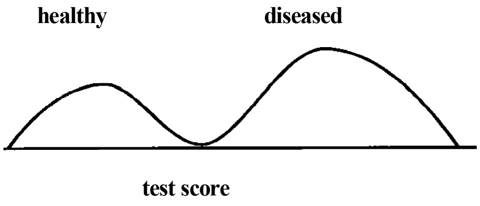
Learning objectives

By the end of this Unit, you will be able to:

- Define the terms *sensitivity*, *specificity*, *positive predictive value*, *negative predictive value*, and describe their importance to health practitioners and patients
- Describe the trade-offs between sensitivity and specificity
- List the factors that contribute to high positive predictive value
- Calculate and interpret sensitivity, specificity, and positive predictive value from sample data.

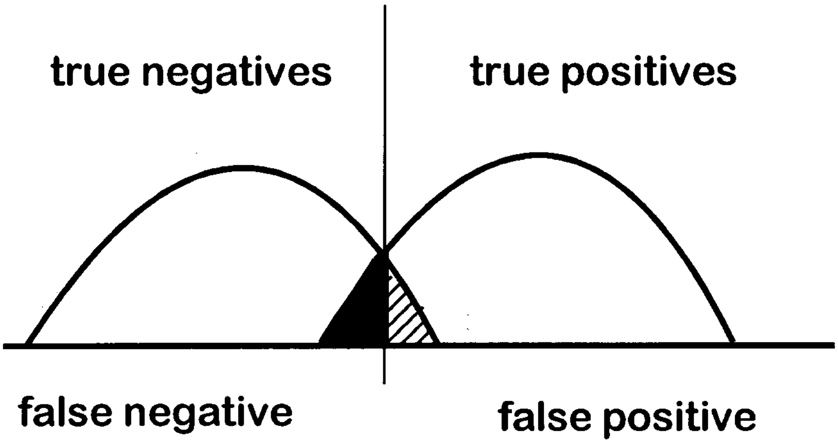
An ideal laboratory test would detect all people who have a disease and at the same time identify as normal all those who do not have the disease.

Figure 9a: Ideal test



However, many laboratory tests (such as haematocrit readings, blood glucose, and optical density testing) are based on continuous data and the values between people with and those without a disease may overlap.

Figure 9b: Practical situation of a test



How well a laboratory test performs in the identification of individuals with or without the disease can be assessed from the values in the following 2 x 2 table:

	Disease present	Disease absent	
Test *positive*	True positives (TP) (a)	False positives (FP) (b)	TOTAL POSITIVE
Test *negative*	False negatives (FN) (c)	True negatives (TN) (d)	TOTAL NEGATIVE

From this table, it is possible to calculate the following values that summarize the performance of the test:

Prevalence	$\frac{\text{People with the disease}}{\text{All people}}$	$\frac{TP + FN}{TN + TP + FN + FP}$	$\frac{(a + c)}{(a + b + c + d)}$
Sensitivity	$\frac{\text{People with the disease and a positive test}}{\text{All people with the disease}}$	$\frac{TP}{TP + FN}$	$\frac{(a)}{(a + c)}$
Specificity	$\frac{\text{People with a negative test without the disease}}{\text{All people without the disease}}$	$\frac{TN}{FP + TN}$	$\frac{(d)}{(b + d)}$
Positive predictive value: % of positives with the disease	$\frac{\text{People with the disease and a positive test}}{\text{All people with a positive test}}$	$\frac{TP}{TP + FP}$	$\frac{(a)}{(a + b)}$
Negative predictive value: % of negative without the disease	$\frac{\text{People without the disease with a negative test}}{\text{All people with a negative test}}$	$\frac{TN}{TN + FN}$	$\frac{(d)}{(c + d)}$

For most laboratory tests, the most critical items are:

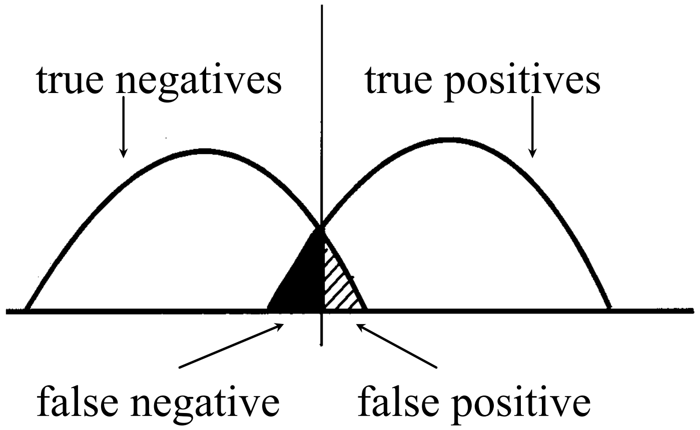
- Sensitivity *(does this test detect as many persons with the disease as possible?)*
- Specificity *(if a test is negative, what is the proportion of false negatives?)*
- Positive predictive value *(among those testing positive, what is the probability that we are dealing with "patients with the disease?)*

For a surveillance system, the most critical items are sensitivity *and* positive predictive value.

You can move down the cutoff point for what constitutes an abnormal value in order to improve the sensitivity of a test but by moving the cutoff point down you increase the rate of false positives. Conversely, if you move the cutoff point upward in order to provide greater specificity, the sensitivity will decrease and you will increase the proportion of tests that are falsely positive. Where you set the cutoff point depends on the consequences of false positives and negatives.

- If you want to reduce the chance of having *false negatives*, as in the case of colon cancer where early action is important, you should set the cutoff point low (high *sensitivity*).
- In some cases, however, a positive text may entail other investigations or treatments that may be risky for the patient (e.g., coronary investigations). In such cases, you should reduce the chance of *false positives* by setting the cutoff point higher for higher *specificity* (even if this affect the *sensitivity* of the test).

Figure 9c: Negatives and positives



Sensitivity and specificity are **independent of prevalence** of a disease. However, positive predictive value is **dependent on prevalence**, as shown by the following:

Example: Suppose you have a new laboratory test that you think will be useful in diagnosing onchocerciasis. You decide to try it out on all patients seen in a clinic who come in with vision problems (their "true" status regarding presence or absence of the disease is subsequently decided using a sophisticated but costly test that is currently being used to make the diagnosis).

After several months of data collection you arrive at the following table:

	Onchocerciasis present	Onchocerciasis Absent	TOTAL
Test Positive	215	16	231
Test Negative	15	114	129
TOTAL	230	130	360

Here:

Prevalence	TP + FN / Total	(215 + 15)/ 360 =	64%
Sensitivity	TP / (TP + FN)	215 / (215 + 15) =	93%
Specificity	TN / (TN + FP)	114 / (114 + 16) =	88%
Positive predictive value	TP / (TP + FP)	215 / (215 + 16) =	93%
Negative predictive value	TN / (TN + FN)	114 / (114 + 15) =	88%

There is an interesting relationship between the *prevalence* of a condition and the *positive predictive value* of the corresponding test. Consider what might have happened if you had screened everyone coming to the outpatient clinic with your new diagnostic test, instead of just those with eye problems. You would have found the following:

	Onchocerciasis Present	Onchocerciasis Absent	TOTAL
Test positive	215	248	463
Test negative	15	1822	1837
TOTAL	230	2070	2300

and

Prevalence	TP + FN / Total	(215 + 15)/ 2300 =	64%
Sensitivity	TP / (TP + FN)	215 / (215 + 15) =	93%
Spécificity	TN / (TN + FP)	1822 / (1822 + 248) =	88%
Positive predictive value	*TP / (TP + FP)*	*215 / (215 + 248) =*	*46%*
Negative predictive value	*TN / (TN + FN)*	*1822 / (1822 + 15) =*	*99%*

Sensitivity and *specificity* remained the same, but the *positive predictive value* and *negative predictive value* changed. Thus, some factors regarding test performance depend on the disease prevalence while others do not. In general – and all other things remaining the same – **the higher the prevalence, the higher the positive predictive value**; conversely**, the lower the prevalence the lower the positive predictive value**.

Tests are sometimes used in sequence in order to maximize their sensitivity and specificity. For example, the ELISA test for HIV has a high sensitivity but a low specificity. In those with a positive ELISA test, this must therefore be followed by a Western blot test, which has higher specificity. Repeating the test on a higher prevalence population also improves the positive predictive value of the test.

Application of sensitivity and specificity to surveillance

In an ideal surveillance system, all cases in the population would be detected, and all those that the surveillance system identified as having the disease would indeed have the disease. In practice, depending on the case definition used, some of those who have the disease will not be included as cases (lack of sensitivity), and some of those that are tested as positive will not have the disease (low specificity). Additionally, not all of those who meet the case definition will actually have the disease (positive predictive value). In addition to problems with the case definition, surveillance systems can have low sensitivity if the surveillance detects only a fraction of the cases that actually occur in the population.

Sensitivity:	$\frac{\text{cases detected by surveillance}}{\text{all people with the disease}}$	$\frac{TP}{TP + FN}$
Specificity:	$\frac{\text{people without the disease and negative on surveillance}}{\text{all people without the disease}}$	$\frac{TN}{TN + FP}$
Positive predictive value:	$\frac{\text{people with the disease detected by surveillance}}{\text{all people meeting the case definition}}$	$\frac{TP}{TP + FP}$

In surveillance, the two most important values are sensitivity and positive predictive value. *Sensitivity* is affected by:

- Whether people with the condition seek medical care
- Whether the disease is diagnosed; and
- Whether the disease is reported.

To evaluate sensitivity, you need external evaluation through a mechanism such as a survey. Recall that sensitivity does not have to be high in order to monitor trends, as long as sensitivity remains relatively constant.

Positive predictive value is important if the surveillance system may trigger the further investigation of individual cases or of outbreaks. If the positive predictive value is low, resources will be wasted chasing problems that do not exist.

Using a broad case definition to improve sensitivity will increase the rate of false positives and decrease specificity. Similarly, increasing the criteria required to make a diagnosis will increase specificity, but sensitivity will decline.

Exercises on sensitivity and specificity

The Minister of Health informs the Director of the malaria programme that there will be a 20% cut in the next yearly budget. The Director looks for ways to cut expenses without compromising the coverage and quality of the malaria programme.

One possibility would be to stop performing routine microscopy on all suspected malaria cases. Microscopy, which is done while the patient waits, has been used for many years in this country to determine who is infected, to distinguish between *P. vivax* and *P. falciparum*, and for surveillance purposes. The *P. falciparum* infections seen in the country are chloroquine-sensitive, so patients with *P. vivax* and *P. falciparum* both receive chloroquine.

A factor influencing the decision whether or not to discontinue microscopy is the rising rate of AIDS in his country, even in rural areas. The Director worries that malaria workers may become infected while doing thick smears or that the infection will be transmitted from inappropriately sterilized lancets or needles.

If he stops microscopy, the programme will have to rely on clinical findings and will be treating some people unnecessarily. To find out how good the clinical signs are at detecting parasitaemia, the Director speaks with a colleague who studied malaria incidence and clinical symptoms in a sample of adults from a high endemicity province (prevalence 2% per week) and from a low endemicity province (prevalence 0.2% per week).. In both areas, 98% of those who develop parasitaemia as diagnosed by a thick smear have fever and rigors, but 1% of those who have negative parasitaemia on thick smear will also have fever and rigors.

Exercise 1

(a) Using the thick smears as the "gold standard" what is the sensitivity of fever and rigors to detect malaria?

(b) What is the specificity of these signs?

	Fever and chills present	*Fever and chills absent*
+ve smear		
-ve smear		
TOTAL		

Exercise 2

For a hypothetical population of 100 000 adults in the **high** endemicity area, calculate the following:

(a) The number of persons who would receive treatment each week, assuming that all those with fever and rigors would be treated.

(b) The number who would be unnecessarily treated.

(c) The positive predictive value.

(d) The negative predictive value.

Exercise 3

State in words the meaning of the positive predictive value and negative predictive value that you have calculated.

The results seem encouraging, and the Director decides to see what would happen in the low endemicity area.

Exercise 4

Results are encouraging and the Director should like to know what would happen in the areas of low prevalence.

For a hypothetical population of 100 000 adults in the **low** endemicity area, calculate the following:

(a) The number of persons who would receive treatment each week, assuming that all those with fever and rigors would be treated.

(b) The number who would be unnecessarily treated.

(c) The positive predictive value.

(d) The negative predictive value.

NOTES

Further reading

If you wish to know more about the subject, you may wish to refer to one or more of the following publications:

Books

BARKER, D.J.P., HALL, A. J. *Practical epidemiology*, Edinburgh: Churchill Livingston, 1991.

BEAGLEHOLE, R., BONITA, R., KJELLSTROM, T. *Basic epidemiology*, Geneva: World Health Organization, 1993.

GIESECKE J. *Modern infectious disease epidemiology*. London: Edward Arnold, 1994.

GREGG M., ed. *Field epidemiology*. New York: Oxford University Press, 1996.

HENNEKENS C.H., BURING J.E. *Epidemiology in medicine*. Boston: Little, Brown, 1987.

LAST, J.M. *A dictionary of epidemiology*, Oxford University Press, 2001.

LWANGA, S.K., TYE, C-Y. *Teaching health statistics*, WHO, 1986.

MAUSNER J.S., KRAMER, S. *Mausner & Bahn Epidemiology—an introductory text*. Philadelphia: Saunders & Co., 1985.

SWINSCOW, T.D.V. *Statistics at square one*, London: BMJ Publishing Group, 1996.

VAUGHAN, J.P., MORROW, R.H. *Manual of Epidemiology for District Health Management*, Geneva: World Health Organization, 1989.

Journals

Major journals such as the *Lancet*, the *British Medical Journal*, the *Journal of the American Medical Association* and the *New England Journal of Medicine* carry items of epidemiological interest. The following deal specifically with epidemiology – you may identify others through local sources.

American Journal of Epidemiology
DESCRIPTION: Occurrence and distribution of endemic and epidemic diseases
FREQUENCY: twice monthly CIRCULATION: c6000
PUBLISHER: Johns Hopkins School of Hygiene and public Health, 615 N. Wolfe Street, Baltimore MD 21205, USA Tel: 001 410 955 3441 Email: sgadams@jhsph.edu

American Journal of Public Health
DESCRIPTION: Current aspects of public health
FREQUENCY: Monthly CIRCULATION: 30000+
PUBLISHER: American Public Health Association, 1015 15th Street NW, Washington DC 20005, USA Tel: 001 202 789 5600
Website: http://www.apha.org

Canadian Journal of Public Health/Revue canadienne de Santé publique
DESCRIPTION: Reports on public health and preventive medicine in Canada
FREQUENCY: 6 times a year CIRCULATION: c3000+
PUBLISHER: Canadian Public Health Association, 1565 Carling Avenue Suite 400 Ottawa, Ontario, KIZ 8R1 Canada Tel: 001 613 725 3769

Epidemiologic reviews
DESCRIPTION: Definitive reviews in epidemiological research
FREQUENCY: twice a year CIRCULATION: c6000
PUBLISHER: Johns Hopkins School of Hygiene and public Health, 615 N. Wolfe Street, Baltimore MD 21205, USA Tel: 001 410 955 3441 Email: sgadams@jhsph.edu & http://phweb.sph.jhu.edu./pub/jepi

Epidemiology and Infection
DESCRIPTION: All aspects, mainly communicable diseases
FREQUENCY: twice a month CIRCULATION: c1300
PUBLISHER: London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK Tel: 0044 71 927 2398
URL: http://www.cup.org/journals/cup.org/

European Journal of Epidemiology
DESCRIPTION: Epidemiology and control of communicable and non-communicable diseases
FREQUENCY: ten times a year CIRCULATION: c1300
PUBLISHER: Kluwer Academic Publisher Spuiboulevard 50 POB 17
3300A Dordrecht, The Netherlands Tel: 0031 78 639 2143 URL: http://www.wkap.nl

International Journal of Epidemiology
DESCRIPTION: Research and teaching, published by the International epidemiological Association
FREQUENCY: six times a year CIRCULATION: c2500
PUBLISHER: Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK Tel: 0044 86 555 6767 Email: p.o.d.pharoah@liv.ac.uk

Journal of Epidemiology and Community Health
DESCRIPTION: Health of communities worldwide
FREQUENCY: four times a year CIRCULATION: —
PUBLISHER: British Medical Journal BMA House, Tavistock Square London WC1 9HR UK Tel: 0044 171 387 4499 Email/URL. http://www.jech.com

Morbidity and Mortality Weekly Report
DESCRIPTION: Outbreaks and control measures, surveillance programmes/projects, guidelines and recommendations on topics of public health concern
FREQUENCY: weekly CIRCULATION: 45 000+
PUBLISHER: Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta GA 30333 Tel: 001404 639 2100; Fax: 001404 639 880; Website: httpo://www.cdc.gov

Pan American Journal of Public Health
DESCRIPTION: Public health in the Western hemisphere
FREQUENCY: monthly CIRCULATION: c6000
PUBLISHER: Pan American Health Organization, 525 Twenty-third Street NW, Washington DC 20037 USA Tel: 001 202 338 0869 Email: sales@paho.org & URL http://www.paho.org

Revue d'Epidémiologie et de Santé publique
DESCRIPTION: Current aspects of public health
FREQUENCY: Monthly CIRCULATION: —
PUBLISHER: Editions Masson 120 Bd Saint Germain, 75280 Paris CEDEX, France Tél : 00331404466200

Weekly Epidemiological Record/Relevé épidémiologique hebdomadaire
DESCRIPTION: Diseases subject to the *International Health Regulations*, reports on major diseases and disease control, epidemiological activities in Member States
FREQUENCY: weekly CIRCULATION: 7300+
PUBLISHER: World Health Organization, Avenue Appia, CH-1211 Geneva 27, Switzerland Tel: 004122 791 2111; Fax: 004122 791 0746
Website: http://www.who.ch/wer/wer_home.htm

Software

DEAN, A.G., DEAN, J.A., COULOMBIER, D., BRADNEL, K.A., SMITH, D.C., BURTON, A.M., DICKER, R.C., SULLIVAN, I., FAGAN R.F., AZNER, T.G. *Epi-Info Version 6.04: a word processing, database and statistics program for epidemiology on microcomputers.* Centers for Disease Control and Prevention, Atlanta, Georgia, USA, 1994.

EPI INFO consists of a series of microcomputer programmes for word processing, data management and epidemiological analysis designed for public health professionals. It also offers programming languages for both data input and analysis towards permanent health information systems.

EPI INFO software contains:

- **Epi Info**: allows rapid set-up of new entry forms and data files, easily customized data entry, and many data management and analysis techniques.

- **Epi Map**: displays counts or rates on geographic maps supplied or drawn on the screen. Colours, shading dots, or noncontiguous cartograms can be used to show any type of numeric data related to map.

- **DoEpi**: a series of educational studies and computer exercises designed to teach both epidemiology and the use of Epi Info. An instructor's module is included.

- **SSS1**: provides functions for Box Jenkins Time Series analysis, "Fig. 1" MMWR graphs, robust trend analysis, and comparison of surveillance data from two sources.

WEBSITES

You can download the programs (in English) from the following:

- CDC Epidemiologic Software (English) http://www.cdc.gov/epo/epi/software.htm
- WHO Homepage WHOSIS (English) http://www.who.int/whosis
- Epi Info Manuals Brixton Books UK (English) http://mkn.co.uk/help/extra/people/Brixton_Books
- USD, Inc. (English): http://www.usd-inc.com/phi.html

For further assistance, contact **Epi Info hotline**
Tel: (00) 1 404 639-0840 Fax: (00) 1 404 639-0841
e-mail: epiinfo@cdc.gov