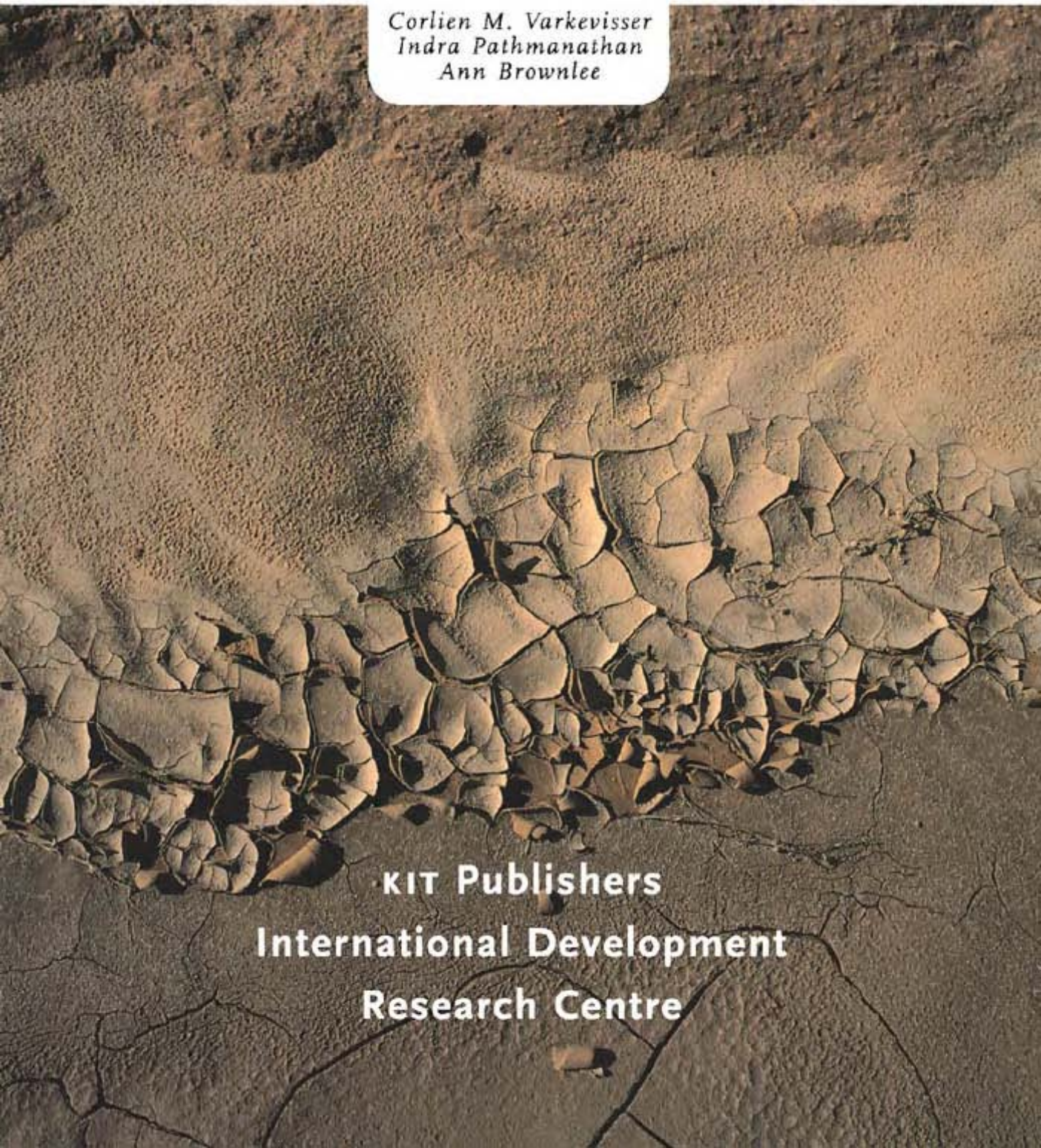
Designing and conducting health system research projects, *volume 2,* Data analyses and report writing

Corlien M. Varkevisser
Indra Pathmanathan
Ann Brownlee



KIT Publishers
International Development Research Centre

Designing and Conducting Health Systems Research Projects

Volume II: Data analysis and report writing

This page intentionally left blank

Designing and Conducting Health Systems Research Projects

Volume II: Data analysis and report writing

Corlien M. Varkevisser
Indra Pathmanathan
Ann Brownlee

KIT Publishers, Amsterdam
International Development Research Centre

in association with
WHO Regional Office for Africa

Designing and conducting health systems research projects
Volume II: Data analysis and report writing

Jointly published by KIT Publishers and the International Development Research Centre (IDRC), in association with the Africa Regional Office (AFRO) of the World Health Organization.

KIT Publishers
Mauritskade 63, 1090 HA Amsterdam, the Netherlands
publishers@kit.nl / www.kit.nl
ISBN 90 6832 148 X

International Development Research Centre
PO Box 8500, Ottawa, ON, Canada K1G 3H9
info@idrc.ca / www.idrc.ca
ISBN 1-55250-069-1 (Volume 1) / 1-55250-070-5 (Volume 2)

World Health Organization - Regional Office for Africa
Cite du Djoue, P.O.Box 06 Brazzaville, Congo
www.whoafr.org

Cover design: Mulder en van Meurs, Amsterdam
Production: Meester & de Jonge, Lochem
NUR 600

TABLE OF CONTENTS

This page intentionally left blank

FOREWORD

Health Systems Research (HSR) has proved to be a useful tool for health decision makers at all levels over the past 20 years, providing them with the necessary data for informed decision making.

The Joint HSR Project for the Southern African Region based in the WHO Office in Harare and supported by WHO Geneva, the Royal Tropical Institute (KIT) in Amsterdam and the Dutch Technical Development Co-operation (DGIS), has played a crucial role in the promotion of HSR in the African region since 1987. HSR was enthusiastically embraced by many Ministries of Health and universities. In 1996, the Regional WHO Office for Sub-Saharan Africa (AFRO) assumed full responsibility for implementing HSR. Following the recommendation of Health Ministers of the Region, WHO/AFRO in 1998 included HSR as a regular programme for all its 46 member states.

The present HSR training modules, developed by an interdisciplinary, international team of practical researchers, have been highly instrumental in raising the interest for HSR. Originally designed for health managers at different levels as a tool to develop problem solving research in the Southern African Region, the modules also proved useful in Malaysia and were further elaborated by staff of the School of Public Health The 1991 combined version, published by International Development Research Centre, Canada and WHO, Geneva,* was translated in French, Spanish and Portuguese, and sections of it appeared in Arabic, Vietnamese and Chinese. In different parts of the world the modules facilitated the development and implementation of hundreds of research protocols by health staff and researchers. The HSRmodules are used in the Community Health and Social Science Departments of many African, Asian and Latin American universities to train students and prepare them for their fieldwork. They are also used by Masters of Public Health courses in Europe and the USA and by international researchprogrammes interested in applied research.

This unanticipated application of the modules in academic as well as health management circles led to the rapid exhaustion of the 1991 edition and the several subsequent reprints. Various groups of users made many useful suggestions for changes and improvements. The HSR Unit in AFRO, with agreement from IDRC, therefore decided to organise a revision of the HSR modules. An interdisciplinary group of Southern African researchers reviewed and made revisions in two workshops in 1998 and 1999. Two of the three original editors finalised the present version. IDRC took on the final responsibility for the publication, which was financed by AFRO and IDRC and published by KIT.

It is hoped that this revised version of the modules will fulfil the same need as preceding ones have done. Certainly many new and persisting health problems urgently require operational research. How to support necessary health reforms and at the same time ensure equity in access to health care for high-risk groups remains a major challenge. HSR is one of the tools we have to obtain deeper insight in these challenges and optimally focus our resources.

Dr. Rufaro R. Chatora, Director of the Division of Health Systems and Services Development (DSD), WHO/AFRO, Harare

Dr. Christina Zarowsky, Senior Health Specialist, IDRC, Ottawa

Ms. Catherine Hodgkin, Head Health Department, KIT, Royal Tropical Institute, Amsterdam

* Corlien M. Varkevisser, Indra Pathmanathan and Ann Brownlee (1991) *Designing and conducting Health Systems Research projects. Part I: Proposal development and fieldword; Part II: Data analysis and report writing.* Ottawa: Health Sciences Division of the International Development Research Centre (IDRC) and Geneva:Programme on Health Systems Research and Development of the World Health Organisation.

This page intentionally left blank

PREFACE AND ACKNOWLEDGEMENTS

The present volume *'Designing and conducting Health Systems Research'* is a thorough revision of Volume 2 of the *Health Systems Training Series* which the International Development Research Centre (IDRC) in Canada and WHO HQ in Geneva published in 1991 and reprinted several times under the same name. It became necessary to revise the modules, because over the years inevitable shortcomings and gaps were detected which needed to be addressed. been added.

Health managers, for example, stressed that implementation of the research findings and recommendations were somewhat underexposed in the modules. This point is now taken care of in Module 1 by adding a fourth, implementation phase to the Health Systems Research (HSR) training cycle which initially consisted of three phases: HSR proposal development (15 days), fieldwork (roughly 6 months) and data analysis and report writing (2 weeks). The implementation of research findings and recommendations is further elaborated in Module 33. Furthermore, health managers pleaded, understandably, for shorter courses. This wish has been taken care of by stressing more explicitly in Modules 1 and 3, as well as in the Course Guidelines (annexed to Part 1 of this volume) that the proposal development phase can be shortened by having research teams select their research topic in the field before the onset of the course, preferably under guidance of a facilitator. In addition, the WHO/AFRO HSR Programme based in Harare, is at present developing modules for participatory rapid action in health research at health centre and district levels which can be carried out and integrated in the day to day activities of staff and community members.

Research staff from Community Health, Social Science and other university departments/ research institutes in Sub-Saharan Africa or other parts of the world who are using the modules had other wishes. They advocated that, in addition to the already well-emphasized problem-solving, analytical research approaches, more weight should be given to descriptive research. A descriptive diagram has therefore been added to the problem analysis diagram in Module 4. In all subsequent research steps, if relevant, the distinction between analytic and descriptive studies has been elaborated. Qualitative research methods have also been given more weight and they were more thoroughly integrated with quantitative methods in the research methodology (Modules 8-14). This applies, for example, to Modules 10 (Data collection techniques) and 11 (Sampling techniques). Furthermore, two new modules have been added to Part 2 (Data analysis and report writing) of the volume: one on Measures of association based on risk (Module 25), which used parts of Module 30 in the 1991 version, and one on the difficult issue of Confounding variables (Module 26). This need for extension was also reflected in the most recent evaluation of HSR training (1997).*

Facilitators, finally, desired more elaborate examples of crucial research and data-analysis techniques. Therefore, Module 10B (Development of research instruments) has been elaborated with a section on interview techniques with interview exercises, and Module 10C (FGDs) now contains an example of a transcribed focus group discussion with codes in the margin. To Module 13 (Plan for data analysis), an example of a full-fledged questionnaire and of a master sheet have been added, and Module 23 (Analysis of qualitative data) now provides an example of a filled-in compilation sheet. Module 5 (Literature review), has been extended with an example of a literature review.

Apart from these additions, in all modules parts that had proven to be unclear or incomplete were rewritten, and many examples and references were replaced by more recent ones or elaborated.

The present revision was initiated in a workshop held from 2-11 November 1998 in Arusha by a group of interdisciplinary researchers and managers convened by the manager of the WHO/AFRO HSR Programme (since 1992 Gabriel Mwaluko). All participants had thorough experience with the modules and with HSR: Sambe Duale, Lawrence Gakuri, Pilate Khulumani, Steve Kinoti, Gabriel Mwaluko, Jude Padayachi, Brian Pazvakavambwa, Corlien Varkevisser and Godfrey Woelk. In August 1999 a group of three people (Alasford Ngwengwe, Corlien Varkevisser and Godfrey

---

* Corlien M. Varkevisser, Indra Pathmanathan and Ann Brownlee (1991) *Designing and conducting Health Systems Research projects. Part I: Proposal development and fieldword; Part II: Data analysis and report writing.* Ottawa: Health Sciences Division of the International Development Research Centre (IDRC) and Geneva:Programme on Health Systems Research and Development of the World Health Organisation.

Woelk) made further revisions and synchronised the different texts in the WHO/AFRO/HSR office in Harare, supported by staff of the HSR office (since 1999 headed by Isabel R. Aleta, with Makhamokha Mohale and Eric Naterop as APOs). Corlien Varkevisser and Ann Brownlee finalised and edited the modules, with the blessing of Indra Pathmanathan who this time could not participate. Deborah Karugonjo (Harare) and Merel Gallée (Amsterdam) provided highly valued assistance in the production of successive computerised versions. Funds for revising and publishing the HSR modules were made available by DGIS (Dutch Development Co-operation); SARA/AED, Washington; GTZ, Germany through the GTZ MCH/FP network for Health Systems Research in Southern Africa; WHO/AFRO and by WHO HQ, Geneva. IDRC, Canada assists in subsidised distribution of the modules.

A highly varied collection of people assisted in the production of earlier versions of the HSR modules. The cradle of the modules stood in Western Africa, where in the early eighties the Project for Strengthening Health Delivery Systems (SHDS), based in Boston University, USA, at the request of AFRO developed training materials in research protocol development. SHDS followed the step-by-step approach which till today is a major key to the success of HSR courses. Modules 1-17 in this volume are heavily adapted or new versions of the original SHDS modules.* The first adaptation took place in 1988, with 12 researchers from countries that participated in the Joint HSR Project (Omondi (Kenya), Sebatane and Makatjane (Lesotho), Chimimba and Msukwa (Malawi), Kitua and Savy (Seychelles), Tembo (Zambia) Munochiveyi, Taylor and Woelk (Zimbabwe) and Joint Project staff which also finalised the version (Corlien Varkevisser and Martien Borgdorff). These 'green modules'* found their way to Malaysia, where Indra Pathmanathan further developed them, with assistance from Maimunah Abdul Hamid, K. Mariappan and C. Sivagnanasundram (Sri Lanka), in the course of numerous protocol development workshops. The same occurred in Southern and Eastern Africa. At the initiative of Yvo Nuyens, who fathered the Joint HSR Project in WHO Geneva, and supported by IDRC (Annette Stark), the five volumes of the *Health Systems Research Training Series* emerged, of which *Designing and Conducting Health Systems Research Projects* formed Volume 2. These 'pink modules', published in 1991 in Ottawa by IDRC and WHO, form a thorough merge of the ever developing Southern African and Malaysian versions. They were integrated in Harare (Corlien Varkevisser and Leon Bijlmakers), in consultation with Indra Pathmanathan, and with thorough editing support from Ann Brownlee, one of the authors of the original SHDS modules. The present HSR modules are therefore a truly global production. It is even impossible to mention every contributor, because many HSR course facilitators and participants through their questions and critical remarks inspired further changes.

With such a colourful and interactive origin it seems highly unlikely that the present reprint will be the last one. Whenever the modules are used, they will be adapted. We hope, however, that in their present form they will last for some years and will be of use to health staff as well as university students.

Dr. Corlien M. Varkevisser, Royal Tropical Institute/University of Amsterdam

Dr. Ann Brownlee, University of California, San Diego

June 2003

---

* Regional Assessment of Health Systems Research Training in Eastern and Southern Africa. HSR Project and SARA/AED,Harare/Washington. SS Ndeki, 1997.

* Ann Brownlee, Thomas Nchinda and Yolanda Mousseau-Gershman (1983) *Health Services Research Course: How to develop proposals and design research to solve priority problems.* Boston: Boston University Health Policy Institute.

* Joint World Health Organisation/Royal Tropical Institute/Dutch Technical Development Co-operation Project on Health Systems Research for the Southern African Region (1988) *Health Systems Research Training Course: How to develop research proposals to solve priority health problems.* Geneva: World Health Organisation. WHO/SHS/HSR/88.3.

INTRODUCTION TO PART II: Data Analysis and Report Writing

This publication is meant to be used in combination with Part I: *Proposal Development and Fieldwork.* Part I consists of 20 training modules which, step-by-step, support course participants in the development of a research proposal and provide useful guidelines for its implementation.

The present Part II, *Data Analysis and Report Writing,* consists of 13 modules. These training modules on data analysis, report writing, and planning for implementation of recommendations, to a still larger extent than those on the development of a research proposal, can be used in a flexible way, depending on:

- the educational level and research experience of the course participants;
- the type of study conducted and type(s) of data collection techniques used; and
- the state in which the data are at the onset of the data analysis and report writing workshop.

If participants have some previous training in research methodology and statistics, and research experience, presentations of modules may be short. In this case the purpose of presenting is mainly to refresh the participants' memories and to guide them towards correct application of appropriate analysis procedures and tests. Some modules may then be combined or shortened.

If participants have neither training nor experience in research, the presentation of the materials in the modules may have to be restricted to the bare essentials required to handle the data that has been collected. Under these circumstances presentations may take longer and should include ample opportunity for asking questions and for classroom exercises.

'Bare essentials' that could be considered include:

- **Module 21** (Orientation to the workshop).
- **All of Modules 22 and 24** (Description of variables and cross tabulation).
- **Module 23** (Analysis of qualitative data, especially **parts I to IV**).
- **Module 25** (Measures of association based on risk: incidence, risk, relative risk and odds ratio). Concentrate on unpaired observations.
- **Briefly: Module 26** (Dealing with confounding variables). What confounding is, and how to deal with it should become clear through some examples.
- **Module 27** (Preparation for statistical analysis: measures of dispersion, normal distribution and sample variation).
- **Module 28** (Choosing a significance test). Concentrate globally on **parts I, II and III**, explaining the rationale for significance tests and how they work, but deal only briefly with part IV, the actual choosing of a significance test, if the groups are not likely to use more than the chi-square and/or the t-test.
- **Module 29** (Determining differences between groups: analysis of unpaired observations). Either the t-test or the chi-square test or both.
- **All of Modules 32** (Report writing) **and 33** (Promoting the dissemination, communication and utilisation of research findings).

Depending on the types of studies that participants have conducted and the analyses their data require, the scope of the presentations can be expanded (more on statistical tests, for example, or more on analysis of qualitative data), or the sequence changed (Module 23 may be presented before module 22 if participants have mainly qualitative data).

- Usually the *first half of the workshop* (one week) is devoted to the finalisation of data processing and to data analysis. All modules related to analysis (21-31) are presented during this week.
- Timing of these presentations has to be done carefully. Modules 21-24 can be presented before data processing has been completed. Modules 25 and 26 could be presented when groups are about to finish data processing and have finished some basic tables. Module 27, preparing for statistical analysis (measures of dispersion, distribution and sample variation) can then follow as well.
- Only when participants are well underway with the preparation of cross-tabulations, should the modules that present various statistical tests be presented.

- The *second half of the workshop* concentrates on report writing, drafting of recommendations, and presentation and discussion in plenary of the main findings and recommendations arising from the studies. In this week there are usually only two presentations: on report writing (Module 32) and on the dissemination, communication and utilisation of research findings (Module 33). The last module is best presented just before the participants draft the summary of findings and the recommendations of their studies.

An example of a schedule for a 2-week course on data analysis and report writing is presented on the following pages.

If the level of participants is high and if the data have been satisfactorily processed before reconvening for the data analysis and report writing workshop, it may be possible to finish the report including a draft Plan of Action within two weeks. Otherwise, the finishing touches will have to be accomplished afterwards. Some support of a facilitator, either life or by computer, will then be required.

EXAMPLE OF A COURSE SCHEDULE

(as used in southern Africa)
Designing and Conducting HSR Projects:
Data Analysis and Report Writing

Date/Time	Session	Responsible Person(s)
Monday		
08.00 – 08.30	Opening remarks	Course Coordinator
08.30 – 09.15	Presentation and discussion of preliminary results	Group 1
09.15 – 10.00	Presentation and discussion of preliminary results	Group 2
10.00 – 10.30	Tea	
10.30 – 11.15	Presentation and discussion of preliminary results	Group 3
11.15 – 12.00	Presentation and discussion of preliminary results	Group 4
12.00 – 12.30	**Module 21: Orientation to the workshop on data analysis and report writing**	Facilitator
12.30 – 14.00	Lunch	
14.00 – 15.30	Group work	
15.30 – 16.00	Tea	
16.00 – 17.00	**Module 22: Description of variables**	Facilitator
17.00 – 18.00	Group work	
Tuesday		
08.00 – 09.00	**Module 23: Analysis of qualitative data**	Facilitator
09.00 – 13.00	Group work (including tea)	
13.00 – 14.00	Lunch	
14.00 – 15.00	**Module 24: Cross-tabulation of quantitative data**	Facilitator
15.00 – 18.00	Group work (including tea)	

Wednesday

08.00 – 13.00	Group work (including tea)	
13.00 – 14.00	Lunch	
14.00 – 15.00	Optional: Presentations of main results of group work: revised objectives, main cross-tables, results of qualitative analysis	All 4 Groups
15.00 – 16.00	**Module 25: Measures of association based on risk** (incidence, risk, relative risk, odds ration)	Facilitator
16.00 – 18.00	Group work (including tea)	

Thursday

08.00 – 09.00	**Module 26: Dealing with confounding**	Facilitator
09.00 – 13.00	Group work (including tea)	
13.00 – 14.00	Lunch	
14.00 – 15.00	**Module 27: Preparing for statistical analysis** (Measures of dispersion, normal distribution and sample variation)	Facilitator
15.00 – 18.00	Group work (including tea)	

Friday

08.00 – 09.00	**Module 28: Choosing a significance test** (20 min) followed by **Module 29, Part I and II (t-test)**	Facilitator
09.00 – 13.00	Group work (including tea)	
13.00 - 14.00	Lunch	
14.00 - 15.00	**Module 29 part III (Chi square test)**	Facilitator
15.00 - 18.00	Group work (including tea)	

Saturday

08.00 – 13.00	Group work (including tea)	
	Modules 30 or 31 (if required)	

Sunday

Free

Monday		
08.00 – 09.00	**Module 32: Report writing**	Facilitator
Rest of day	Group work	
Tuesday		
Whole day	Group work	
Wednesday		
08.00 – 08.30	**Module 33: Promoting the dissemination, communication and utilisation of findings**	Facilitator
Rest of day	Group work	
Thursday		
08.00 – 13.00	Group work	
13.00 – 14.00	Lunch	
14.00 – 17.30+	Group work; Preparation of presentation of research results, recommendations, and preliminary Plan of Action	All groups and facilitators
Friday		
08.00 – 13.00	Group work to finalise reports and presentations	
13.00 – 14.00	**Lunch, with invited guests** (Interested health and research managers of the MOH and university/research institutions)	
14.00 – 17.00	**Presentations of summary of findings, recommendations and tentative Plan of Action** by the four groups, followed by discussion (3/4 hour per group, with tea break in between)	
17.00 – 17.30	**Evaluation** of the HSR Training Course	
Saturday (if necessary)		
08.00 – 13.00	Finishing touches to reports	

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 21

ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

Module 21: ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

OBJECTIVES OF THE WORKSHOP

At the end of this workshop you should be able to:

1. **Identify** and define the basic concepts and procedures required for data analysis and interpretation.
2. **Analyse** and interpret the data collected for the research project which you developed during the first workshop and draw conclusions related to the objectives of your study.
3. **Write** a clear and concise research report and a summary of the major findings and recommendations for each of the different parties interested in the results.
4. **Present** the major findings and the recommendations of your study to policy-makers managers and to the subjects of your research together with them to finalise the recommendations.
5. **Prepare** a plan of action for the dissemination, communication and utilisation of the findings and (if required) make recommendations for additional research.

I. Review of the field experience

II. Introduction to the workshop

III. Tasks to be completed during the workshop

1. **Review and finalisation of data processing**
2. **Data analysis**
3. **Report writing**
4. **Presentation of summary of findings and recommendations**
5. **Drafting a plan for the implementation of the research results**

I. REVIEW OF FIELD EXPERIENCES

Implementing your planned project proposal must have been a big challenge to you. No doubt you met a number of unexpected obstacles as you became involved in your fieldwork but you will have experienced successes as well. If everything went according to plan you have collected your data; you have processed a large part of it, if not all; you have completed some of your analysis; and you have written a preliminary report on the experiences and results of your fieldwork. Your practical experiences in conducting the project are invaluable. Sharing these experiences with others in this workshop will be a very useful exercise, as both your problems and successes can provide valuable lessons for the future.

Before providing an overview of the focus and activities of this workshop, we would like to spend some time listening to the experiences of each of the research groups.

EXERCISE: Presentations of field experiences

Present the preliminary report that your group prepared at the end of your field experience, following the guidelines given in **Module 20**. Be prepared to answer any questions other participants or facilitators may pose at the end of your presentation.

Each group will have approximately 10 – 15 minutes for its presentation and then some time for questions and discussion.

II. INTRODUCTION TO THIS WORKSHOP

This workshop is a follow-up of the workshop in which you developed your proposal. Now you have the major task ahead of fully analysing the data you brought with you from the field and writing your research report. The report should contain feasible and useful recommendations, based on the findings of your study concerning how to solve the problem investigated.

As in the first workshop there will be presentations, group work sessions and a few plenaries. In this workshop, however, group work will take up most of the time. The presentations will be concentrated in the first week, which will be devoted to data analysis. The second week will be fully reserved for report writing, with only two presentations to guide you. Toward the end of that week an important plenary is planned in which each group will present a summary of its main findings and recommendations. A selected group of policy-makers and health managers who requested the study or have a direct interest in the topic and some interested researchers will be invited to comment on your presentation during that plenary.

The modules for this workshop cover several major tasks, which are schematically presented in the diagram on the next page of this module. This diagram is presented again at the beginning of each subsequent module, to indicate which task is the focus of the presentation. We will now briefly introduce each of these tasks.

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams.

I. TASKS TO BE COMPLETED DURING THE WORKSHOP

1. Review and finalisation of data processing

Although we trust that all groups have put great effort in the processing of their data, some adjustment and elaboration may be required. This is normal in research: each new step forward may be followed by half a step backward. Even in an advanced stage of data analysis you may still have to regroup and reprocess some of your data.

Before beginning data analysis, it is extremely important to check whether data processing has been carried out in such a way that information:

- is easy to handle; and
- has been checked for mistakes that may have crept in during data collection.

You therefore have to ask yourself the following questions:

- **Have the data been sorted appropriately?** Have questionnaires and checklists been numbered in the most convenient way? Can major categories of informants (e.g., m/f, cases and controls) be clearly distinguished to facilitate comparison on relevant variables, as required by your research objectives?

- **Have quality checks been performed** on all data for completeness and consistency of information? Look at **Module 13** for measures to be taken in case of incompleteness and/or inconsistencies.

- **Have all data been entered in the computer or, if using master sheets, have all data been filled?** Do the total number of responses match with the total number of respondents for each variable? If not: have some unknowns or missing data been overlooked?

- **Has all qualitative data been categorised** as far as possible? If applicable, has coding been completed? (See **Module 13** for post-categorising of open-ended questions.) Have FGDs been carefully read and ordered according to the discussion topics? Have particularly illustrative parts in relation to the research objectives or research questions been highlighted? Since for qualitative data the collection, ordering, summarising and analysis are, in principle, intertwined (see module 10C), you will already have coded and interpreted a large part of your qualitative data. **Module 23** will take the analysis of qualitative data up in detail.

- If you used the computer to process your data, **check the frequency counts for each variable** in the questionnaire. Also check the computer cross-tabulations. Details on how to do this are given in **Annex 21.1**.

Before reviewing the data processing procedures it is strongly advised that you make an INVENTORY of all data available for each OBJECTIVE. This is especially important if the data required has been collected using different data collection tools.

Example:

Data sources for Objective 3: 'Detection of weaknesses in the functioning of MCH services, explaining low utilisation of delivery care:'

- Questionnaire for mothers, Questions 12, 15 - 19, 23
- Focus group discussion with health staff, topics 3 and 4
- Observations included in checklists

Such an inventory will help you to better organise data analysis and, later, report writing.

2. Data analysis

When beginning data analysis, we should consider which of our data are quantitative and which are qualitative.

Quantitative data

Quantitative data are expressed in numbers and they are usually presented in frequency tables. From your data master sheets you can easily derive totals for each variable/question, count the number of different answers obtained and present the information in frequency tables. (See **Module 22.**) When analysing quantitative data it is important to consider the aim of your study. Is it to:

- **Describe variables?**

  For example: the distribution of teenage pregnancies in a certain population

- **Look for differences between groups?**

  For example: differences between old settlers and newcomers in a certain area, with respect to income or health status

- **Determine associations between variables?**

  For example: the association between work satisfaction of nurses and the number of staff meetings over the past year

Cross-tabulations are an important tool to summarise and analyse these data (**Module 24**), though there are other possibilities (see **Modules 22, 25** and **31**).

After frequency distributions and different types of cross-tabulations have been made, the type of statistical analysis required has to be selected in order to determine whether the differences and associations found are significant or just a consequence of chance. The selection of appropriate significance tests is elaborated in **Module 28**. In **Modules 29-31** some more advanced statistical concepts for the analysis of quantitative data will be introduced.

The most common significance tests are:

- Student's t-test and the chi-square test to determine differences between groups if observations are unpaired (**Module 29**).

- The paired t-test and McNemar's χ^2 (chi-square) test to determine differences between groups for paired observations (**Module 30**).

For measuring associations between variables, the concepts of Odds Ratio (**Module 26**) and Regression and Correlation (**Module 31**) will be introduced.

Throughout the process of data analysis it is important to keep in mind that *our findings should provide answers to our research questions and thus meet our research objectives. We will eventually want to draw conclusions and make recommendations for action, based on these findings.*

Qualitative data

You will remember that we may obtain qualitative data through:

- **open-ended questions**, not precategorised, in questionnaires or interview schedules which also collect quantifiable data;
- **loosely structured interviews** with predominantly open-ended questions, directed at key informants (individuals) or small groups;
- **focus group discussions** on selected issues, with lists of points to guide the discussions;
- **observations** describing individual or group behaviour;
- **diaries, essays**, and any information that originates from **projective methods** (e.g., unfinished sentences, stories with a gap, free associations of informants with pictures or films shown).

As you will remember from the exercise you did in **Module 13**, the answers to open-ended questions may be:

- **listed**
- **categorised** (based on your research objectives and common sense, combining the answers that belong together in some 4 to 6 categories, rarely more)
- **coded/ labelled**
- **interpreted** per category for content, depending for what purpose you need the data
- **inserted**, using these codes, **in your master sheets, or in the computer**
- **counted**, like other quantitative data

You may have already processed some clear-cut open questions. We will discuss these procedures again in-depth in **Module 23** in case you experienced some problems.

Note:

The major characteristic of analysis of qualitative data is that we analyse it **IN WORDS**, rather than in numbers.

Qualitative data from other sources than open-ended questions require more elaborate coding and compilation techniques. Often it is useful to summarise qualitative data in compilation sheets, diagrams, flow-charts, or matrices which help us in our analysis. **Module 23** will deal with the analysis of such qualitative data in more detail.

3. Report writing

You will be expected to go home with a completed report of your research. It will have the following components:

1. An INTRODUCTION, covering the statement of the problem, some relevant contextual data and literature review.

2. OBJECTIVES

3. A METHODOLOGY section with information on when, where and how you have collected your data, how you have analysed the data; and possible weaknesses in the collection and analysis.

4. FINDINGS

5. DISCUSSION

6. CONCLUSIONS AND RECOMMENDATIONS

The last three sections, which will form the bulk of your report, will be discussed in detail in **Module 32**. The first three sections can be revised and summarised from the relevant sections in your research proposal.

4. Presentation of summary of findings and recommendations

Since an important goal of your research is that appropriate action will be undertaken based on the results of your study, it is important that all parties concerned get an opportunity to discuss findings and recommendations before the report is finalised. You may wish to include policy-makers, health managers, staff and community members or even the media in such a discussion. **Module 33** provides some guidelines on how to organise meetings for this purpose.

5. Drafting a plan for the implementation of the research results

You drafted a plan for utilisation and dissemination of results during the previous workshop (**Module 17**). At the end of the present workshop this plan will be reviewed and developed in more detail, including all parties concerned in the planning for implementation of the recommendations that resulted from your study (**Module 33**).

GROUP WORK (Time flexible, depending on the research topics and state of data processing)

- Reconsider the objectives from your research proposal and **list the different data sources for each objective** (questionnaires, records, focus group discussions etc.).

  NB: If you discover that you have collected more data to explain your research problem than your objectives require, you may review your objectives or add one or two additional ones. However, if you collected less data, don't drop objectives which you could not meet but explain why you couldn't.

- Verify whether all data has been checked for completeness, consistency, and proper coding. If not, do so.

- Determine whether different master sheets have been prepared for different study populations or for different categories of informants you would like to compare to each other. This will facilitate analysis.

  You may also mark the questionnaires of sub-groups with different colours so that you can easily refer back to the raw data to check certain questions.

- Check whether the master sheets have been completed and whether the number of responses for each variable agrees with the number of respondents.

- Determine whether all data that should have been entered in the computer have indeed been entered and cleaned. (See Annex **21.1.**)

- Check whether qualitative data were categorised and summarised in the field. If not, read and order the data, putting discussion topic numbers and additional key words and comments in the margins. In **Module 23** further advice will be provided.

- As you list your different data sources by objective, check whether you have recorded all your relevant observations.

Annex 21.1: Computer output

The hard copy printed out by the computer is the result of the commands used in the computer programmes to analyse the available data. The accuracy of the information printed out is therefore dependent on:

- the data that was entered
- the programmes that were used

The saying, 'garbage in, garbage out', is very apt for computer processing. It is the combined responsibility of the research team and computer specialist to ensure that the information printed out is accurate.

Types of computer printouts

1. List of data

This is a list of the data that was entered into the computer. This printout is helpful if you need to make corrections on the existing data while in the process of validating it.

2. Frequency count

This gives a count (and percentage) of each variable in the questionnaire. See the sample given of a frequency count and note how it relates to the questionnaires.

To ensure that the programmes are correct, the computer specialist must be familiar with the format of the questionnaire used and the process of data collection.

For example:

Was stay in hospital appropriate? (n = 3306)

Yes	No	Don't know	Total
1937	1369	0	3306

Was admission appropriate? (n = 3306)

Yes	No	Don't know	Total
634	719	16	1369

In this example, the total responses for 'yes', 'no' and 'don't know' for the question 'Was the stay in hospital appropriate?' is equal to the total study cases. The total responses for the question, 'Was admission appropriate' are only 1369. The computer specialist must be aware that this question was only asked to the informants who were dissatisfied with their stay in hospital, otherwise the computer print will indicate that 1937 answers are missing instead of not applicable (NA).

A frequency count should be obtained for every question in the questionnaire. Use the frequency count to ensure that:

- the total number of responses in each question is correct (i.e., it should tally with the sample size of persons being asked the question);
- all codes are relevant to the question. **For example**, there should be no codes 3-8 in a question that has only two possible responses (e.g. sex: M or F) and a code for 'unknown' (unknown is usually given a code 9).

3. Cross tabulation

The next commonest computer output is a **cross-tabulation.** This is a table showing the number of subjects who have two (or more) of the variables studied.

Example:

	Male	Female	Total
ill			
not ill			
total			

Before using it, check the cross-tabulation for the following:

- The grand total in the table should correspond to the number of subjects in the sample
- Column and row totals should correspond to the frequency counts for each variable (i.e., the number of males and females should correspond to the respective frequency counts)
- Similarly, numbers 'ill' and 'not ill' should correspond to that frequency count. If these do not correspond, there is probably an error in the programme. Consult your computer specialist.
- If there is a statement in the computer printout showing 'missing cases' it means either:
  - there is a wrong code in the data entry (e.g., code 4 when only 1,2 or 9 is possible), or
  - the categories you have specified are not comprehensive.

  For example:
  The questionnaire allowed for 'unknown' but the computer programme did not. Therefore all cases 'unknown' would appear as 'missing cases'.

  Marital status in the questionnaire allowed for 'married, single, divorced, widowed'. However, the computer programme specified only 'married, single, divorced'. All widowed persons would be missing.

  If the age categories are 10 to 14, 15 to 19 but the programmer accidentally programmed the categories as 10 to 13, 15 to 19, all subjects aged 14 would be missing.

Trainer's Notes

Module 21: ORIENTATION TO THE WORKSHOP ON DATA ANALYSIS AND REPORT WRITING

Timing and teaching methods

3 hours	Presentations of field experiences
3/4 hour	Introduction and discussion
1 hour +	Group work (duration depending on research topics and state of data processing)

Introduction and discussion

- Spend the first part of the introductory session on the participants' reports of their field experiences. If all the groups have come prepared to present their preliminary reports, the session can begin directly with this activity. However, if groups still need some time to prepare for their presentations, time should be arranged for this, either before or at the beginning of this first session.

- The introduction to the workshop should clearly stress that there are different tasks to complete, of which data analysis and reporting of the findings will be the most time consuming. It should be clear to the participants, however, that the preparation of recommendations and their implementation is the ultimate aim of their research projects. You might ask the participants for *suggestions* concerning *which policy makers and managers should be invited* for the presentation and discussion of their research findings and recommendations at the end of the workshop.

- When presenting the diagram consider using different, overlapping transparency sheets.

- Adjust the presentation to the level and interests of the participants. Refresh their memory with examples of the processing of open-ended questions and by explaining the difference between descriptive studies, comparison of groups, studies looking for differences between groups, and studies determining associations between variables - **preferably with examples from their own research.**

- Do not frighten groups that have little statistical experience with details on tests at this stage. Merely state that each type of study requires different tests.

- Stress the importance of listing all the data available for each objective, including qualitative data. As the workshop proceeds there will be so much emphasis on the preparation of tables that participants will tend to forget valuable observations and information obtained from key informants. The facilitator should ask the participants to record this information now (if not already done) and include it in the list of data available for each objective. Check, when the report is being written, that it has been analysed.

Group work

- Read the group work directions along with the group members. Let them re-examine their objectives, list the data available for each objective, and discuss whether the objectives are specific enough to cover all relevant data collected. Sometimes objectives have to be split up, rephrased, added, or their order changed to facilitate analysis. **Never** allow the group to omit an objective without an explanation (in the methodology section) concerning why it could not be met.

- Examine, with the group members, all available data for completeness, mistakes, etc. Make sure that separate master sheets have been prepared for different study populations or for different subgroups that will be compared, or that the data on different subgroups can be easily retrieved from the computer.

Take extra time, as a facilitator, to internalise all data available, to identify possible weaknesses, and to consider various possibilities for analysis. Unless you do this at the onset of the workshop it will be difficult to guide the groups efficiently so that they will obtain optimal results from the data they have collected.

- Group members may work in sub-groups to finalise the data processing, but make sure that you discuss problems and progress with the group as a whole at regular intervals.

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 22

DESCRIPTION OF VARIABLES

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 22: DESCRIPTION OF VARIABLES

OBJECTIVES

At the end of this session you should be able to:

1. **Describe** data in terms of frequency distributions, percentages, and proportions.
2. **Use** figures to present data.
3. **Explain** the difference between mean, median and mode.
4. **Calculate** the frequencies, percentages, proportions, ratios, rates, means, medians, and modes for the major variables in your study that require such calculations.
5. **Identify** other independent variables (in addition to the ones identified during the first workshops), if any, that are necessary in the analysis of your data.

I . Introduction

II. Frequency distributions

III. Percentages, proportions, ratios, and rates

IV. Figures

V. Measures of central tendency

I. INTRODUCTION

When you selected the variables for your study in **Module 8**, you did so with the assumption that they either would help to define your problem (dependent variables) and its different components or that they were contributory factors to your problem (independent variables). The purpose of data analysis is to identify whether these assumptions were correct or not, and to highlight possible new views on the problem under study. **The ultimate purpose of analysis is to answer the research questions outlined in the objectives** with your data.

First, before we look at how variables may be affecting one another, we need to summarise the information obtained on each variable in simple tabular form or in a figure.

Some of the variables may have produced numerical data, while other variables produced categorical data. In analysing our data, it is important first of all to determine the type of data that we are dealing with. This is crucial because the type of data used largely determines the type of statistical techniques that should be used to test whether the results of the study are significant.

Categorical data

There are two types of categorical data: they are nominal or ordinal (see **Module 8**).

In NOMINAL DATA, the variables are divided into a number of named categories. These categories, however, cannot be ordered one above another (as they are not greater or lesser than each other).

For example:

NOMINAL DATA	CATEGORIES
Sex	male, female
Marital status	single, married, widowed, separated/divorced

In ORDINAL DATA, the variables are also divided into a number of categories, but these can be ordered one above another, from lowest to highest or vice versa

For example:

ORDINAL DATA	CATEGORIES
Level of knowledge	good, average, poor
Opinion on a statement	fully agree, agree, doubt, disagree, totally disagree

Numerical data

We speak of NUMERICAL DATA if they are expressed in numbers

There are two types of numerical data: they are discrete or continuous.

DISCRETE DATA are a distinct series of numbers.

For example:

DISCRETE DATA	VALUES
Number of motor vehicle accidents	0, 1, 2, 6, 19, etc.
Number of clinic visits	2, 4, 10, 0, 3, etc.
Number of pregnancies per woman	2, 12, 5, 0, 5, 4, etc.

CONTINUOUS DATA come from variables that can be measured with greater precision, depending on the accuracy of the measuring instrument, and each value can increase or decrease without limit.

For example:

CONTINUOUS DATA	VALUES
Height (to 2 decimal point)	12.12, 9.95, 45.13, 6.99, 28.78, etc,
Temperature (in degree Celsius)	37.5, 37.8, 39.2, 40.1, 36.9, etc,
Age (at the last birthday)	50, 45, 12, 78, 25, 16, 61, 90

Numerical data can be presented as:

- Frequency distributions
- Percentages, proportions, ratios and rates
- Figures
- Measures of central tendency

We will now discuss these operations one after each other for both categorical and numerical data.

II. FREQUENCY DISTRIBUTIONS

A FREQUENCY DISTRIBUTION is a description of data presented in tabular form so the data will be more manageable. It gives the frequency with which a particular value appears in the data.

In your research project you will have done already straight frequency counts for all variables in your data master sheets by counting the number of responses in each category. We will now briefly summarise some important points.

1. Categorical data may have very simple categories.

Example 1:

To identify what family planning methods were used by teenagers in Kweneng, West Botswana, teenagers were asked what method they were using.

The results are presented in the following **frequency distribution:**

Method	Number
Abstinence	14
Condoms	47
Injectables	1
Norplant	1
Pill	35
None	307
Total	405

These data are **NOMINAL**. A frequency distribution is calculated by simply totalling the number of responses in each category.

You should always check that the total number of responses agrees or tallies with the number of subjects (respondents). If necessary, there should be a category for missing answers.

We usually express frequency distributions in percentages (see Part III of this Module). By looking at the frequency distribution above you can conclude that roughly 75% or three out of four of the teenagers are not using family planning. For those who are using family planning methods, condoms and pills are the most commonly used methods.

Example 2:

Health personnel from 148 different rural health institutions were asked the following question: How often have you run out of drugs for the treatment of malaria in the past two years? This was a closed question with the following possible answers: never, 1 to 2 times (rarely), 3 to 5 times (occasionally), more than 5 times (frequently).

The number of responses in each category was totalled to give the following frequency distribution:

Categories	Number
Never	47
Rarely	71
Occasionally	24
Frequently	6
Total	148

In this example, the data are **ORDINAL**. The ordering of the categories is important as each category from top to bottom indicates increasing severity of the problem.

The frequency distribution results indicate that most clinics rarely experience shortages of anti-malarial drugs, but that it is an occasional problem in about one sixth of the clinics and a severe problem in a few.

2. Numerical data

Procedures for making frequency distributions of numerical data are very similar to those for categorical data, except that now the data have to be grouped in categories. The steps involved in making a frequency distribution are as follows:

1. Select groups for grouping the data.
2. Count the number of measurements in each group.
3. Add up and check the results.

When grouping data, the way the groups are selected can affect what the results are going to look like. There is little substitute for common sense here, but it may be necessary to change the grouping if you suspect the information is being hidden by a poor selection of the groups.

Example 3:

Health centres of District X are submitting numbers of malaria cases and you wish to summarise them. Compare the daily and weekly summaries of the same data as presented in **Table 22.1**:

Both daily and weekly data show an increasing amount of malaria, but the improving situation shown in days 19, 20 and 21 is not reflected in the weekly summary. It would therefore be better to use the daily data if you want to indicate when exactly the numbers of reported malaria cases started going down.

Table 22.1: Daily and weekly summaries of malaria cases in health centres in District X

Day 1	9 cases		
Day 2	12		
Day 3	11		
Day 4	13		
Day 5	14		
Day 6	13		
Day 7	16	Week 1	88 cases
Day 8	16 cases		
Day 9	16		
Day 10	18		
Day 11	19		
Day 12	16		
Day 13	21		
Day 14	25	Week 2	131 cases
Day 15	28 cases		
Day 16	28		
Day 17	28		
Day 18	32		
Day 19	21		
Day 20	19		
Day 21	12	Week 3	168 cases

When grouping data the following rules are important:

- The groups must not overlap, otherwise there is confusion concerning in which group a measurement belongs.

- There must be continuity from one group to the next, which means that there must be no gaps. Otherwise some measurements may not fit in a group.

- The groups must range from the lowest measurement to the highest measurement so that all of the measurements have a group to which they can be assigned.

- The groups should normally be of an equal width, so that the counts in different groups can easily be compared.

Sometimes, however, it is valid to choose groups that are of different widths, for example if you are interested in specific age groups (e.g., less than 1 year, 1 to 4 years, 5 to 14 years).

When you start summarising data it is better to make too many groups than too few. This is because during data analysis you can combine groups to form new categories without having to go through all your data again, whereas if you have too few groups you have to go back to your raw data to make more groups.

A larger number of groups will generally give a more precise picture, but when using too many groups one can lose the overview.

As a general rule choose round numbers for the lower values of the group limits.

For example: 1.00-9.99, 10.00-19.99, 20.00-29.99, or:
0-4; 5-9, 10-14, etc.

III. PERCENTAGES, PROPORTIONS, RATIOS, AND RATES

1. Percentages

Instead of presenting data in frequency tables using absolute numbers it is often better to calculate percentages.

A PERCENTAGE is the number of units in the sample with a certain characteristic, divided by the total number of units in the sample and multiplied by 100.

Percentages may also be called RELATIVE FREQUENCIES. Percentages *standardise* the data, which means that they make it easier to compare them with similar data obtained in another sample of different size or origin.

Example 4:

82 clinics in one district were asked to submit the number of patients treated for malaria in one month. The researchers presented both the frequency distribution and percentages (or relative frequencies):

Table 22.2: Distribution of clinics according to number of patients treated for malaria in one month

Number of patients	Number of clinics	Relative frequency
0 to 19	25	31%
20 to 39	3	4%
40 to 59	5	6%
60 to 79	11	14%
80 to 99	19	24%
100 to 119	10	12%
120 to 139	4	5%
140 to 159	3	4%
Total	80	100%

Note:

Usually you do not include missing data in the calculation of percentages.

The frequency of responses in each group is calculated as the percentage of those study elements for which you obtained data (or, if a question is being asked to interviewees, the percentage of those interviewees who answered the question).

However, the number of missing data (e.g., people who did not respond to a question) is a useful indication of the adequacy of your data collection. Therefore this number should be mentioned, for example as a note to your table. (See **Table 22.2**.)

Remember that 'don't know' is a special category that should NOT be counted as missing data. If applicable, 'don't know' should appear as a category in the table.

One should be cautious when calculating and interpreting percentages if the total number is small, because one unit more or less would make a big difference in terms of percentages. As a general rule, percentages should not be used when the total is less than 30.

Therefore it is recommended that the number of observations or total cases studied should always be given together with the percentage.

2. Proportions

Sometimes relative frequencies are expressed in proportions instead of percentages.

A PROPORTION is a numerical expression that compares one part of the study units to the whole; A proportion can be expressed as a FRACTION or in DECIMALS.

Example 5:

Out of a total of 55 patients attending a clinic on a specific day 22 are males and 33 are females. We may say that the proportion of males is 22/55 or 2/5, which is equivalent to 0.40.

Note that when a proportion expressed in decimals is multiplied by 100, the value obtained is a percentage. In the example, 0.40 is equivalent to 40%.

3. Ratios

A RATIO is a numerical expression that indicates the relationship in quantity, amount or size between two or more parts.

In **Example 5** above the ratio of males to females is 22:33, or 2:3.

4. Rates

A RATE is the quantity, amount or degree of a disease or event measured over a specified period of time

Commonly used rates in the health sector are:

- Birth Rate = The number of live births per 1000 population over a period of one year
- Death Rate = The number of deaths per 1000 population over a period of one year

- Infant Mortality Rate (IMR) = The number of deaths of infants under one year deaths of age per 1000 live births over a period of one year
- Maternal Mortality Rate (MMR) = The number of maternal pregnancy-related in one year per 100,000 total births in the same year
- Incidence Rate = The number of new cases per population over a specific period of time (usually a year)
- Prevalence Rate = The number of existing cases per population over a specific period of time (usually a year)

IV. FIGURES

If your report contains many descriptive tables, it may be more readable if you present the most important ones in figures.

The most frequently used figures for presenting data include:

- Bar charts
- Pie charts

} for categorical data

- Histograms
- Line graphs
- Scatter diagrams
- Maps

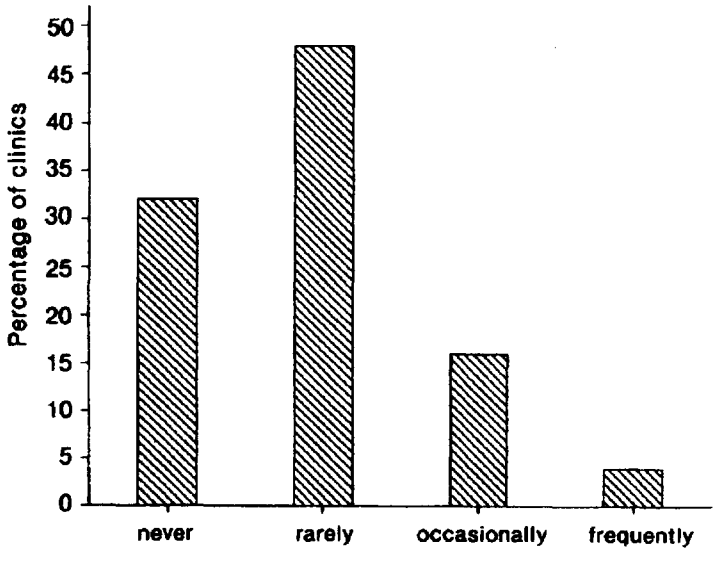
} for numerical data

We will now look at example of the above-mentioned figures that can be used for presenting data.

1. Bar chart

The data from **Example 2** can be presented in a bar chart, using either absolute frequencies or relative frequencies/percentages (see **Figure 22.1**).

Figure 22.1: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n = 148)

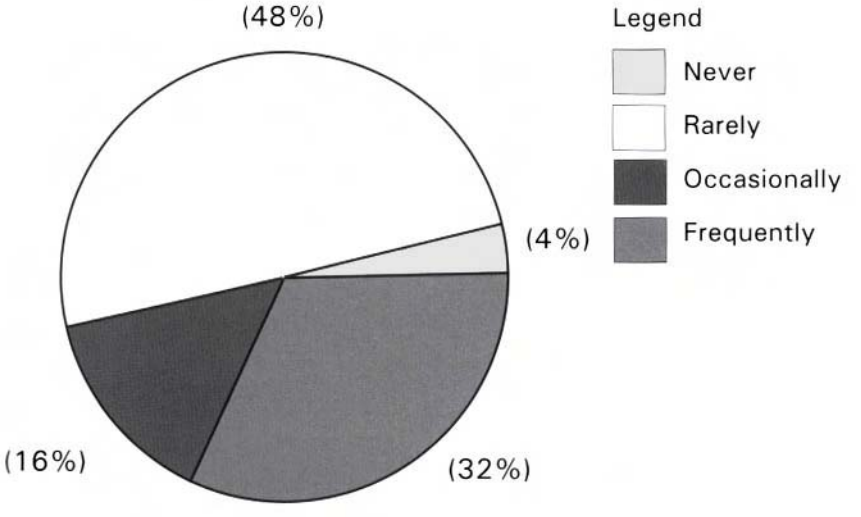


Note that the sample size must be indicated if you present the data in percentages.

2. Pie charts

A pie chart can be used for the same set of data, providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items. All the segments of the pie chart should add up to 100%.

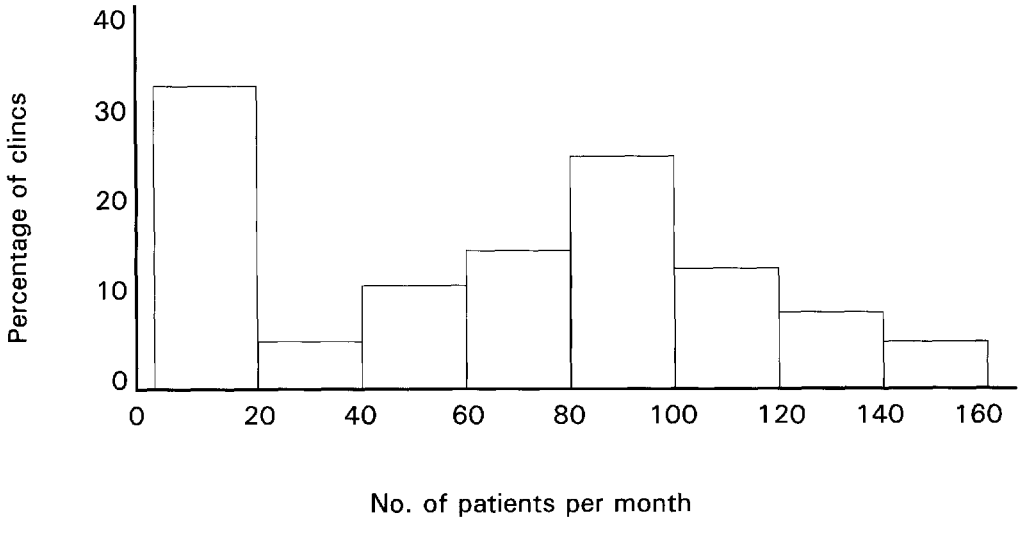
Figure 22.2: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n = 148)



3. Histograms

Numerical data are often presented in histograms, which are very similar to the bar charts which are used for categorical data. An important difference however is that in a histogram the 'bars' are connected (as long as there is no gap between the data), whereas in a bar chart the bars are not connected, as the different categories are distinct entitles. The data of **Example 4** is presented as a histogram in **Figure 22.3.**

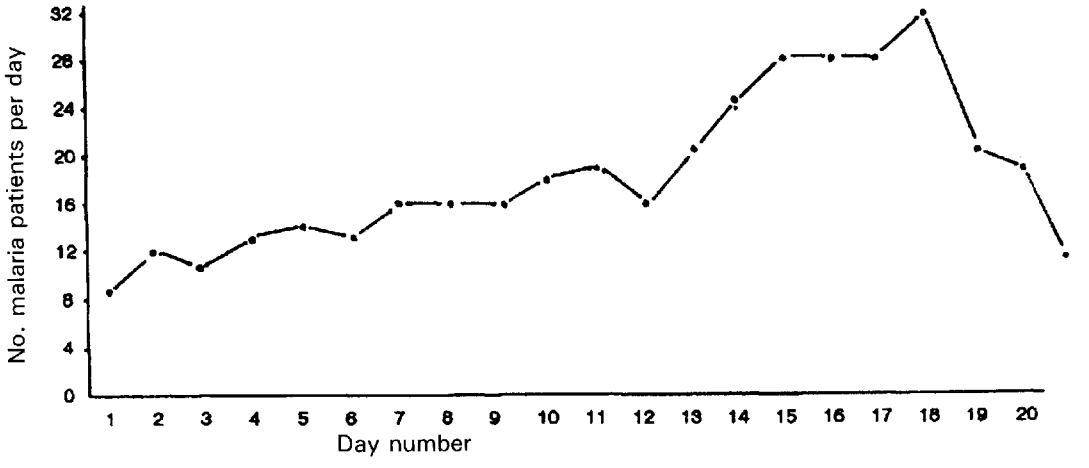
Figure 22.3: Percentage of clinics treating different numbers of malaria patients in one month (n = 80).



4. Line graphs

A line graph is particularly useful for numerical data if you wish to show a trend over time. The data from **Example 3** can be presented as a line graph as in **Figure 22.4.**

Figure 22.4: Daily number of malaria patients at the health centres in District X

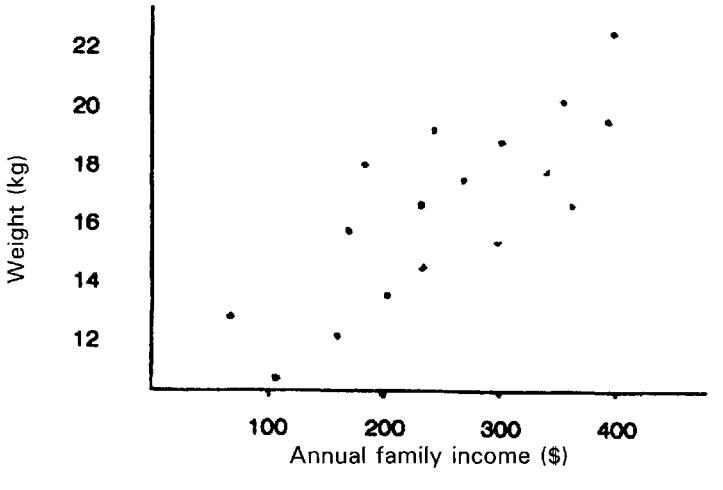


It is easy to show two or more distributions in one graph, as long as the difference between the lines is easy to distinguish. Thus it is possible to compare frequency distributions of different groups, i.e., the age distribution between males and females, or cases and controls.

5. Scatter diagrams

Scatter diagrams are useful for showing information on two variables which are possibly related. The example of a scatter diagram given below is used in **Module 31**, where we are dealing with the concepts of association and correlation.

Figure 22.5: Weight of five-year-olds according to annual family income



Note:

It is important that all figures presented in your research report have numbers, clear titles and clear labels (or keys).

In addition to the figures above, the use of **maps** may be considered to present information. For instance, the area where a study was carried out can be shown in a map. If the study explored the epidemiology of cholera, a map could be produced showing the geographical distribution of cholera cases, together with the distribution of protected water sources, thus illustrating that there is an association. If the study related to vaccination coverage, a map could be developed to indicate the clinic sites and the vaccination coverage among under-fives in each village, perhaps showing that home-clinic distance is an important factor associated with vaccination status.

V. MEASURES OF CENTRAL TENDENCY

Frequency distributions and histograms provide useful ways of looking at a set of observations of a variable. In many circumstances, it is essential to produce them to understand the patterns in the data. However, if one wants to further summarise a set of observations, it is often helpful to use a measure which can be expressed in a single number.

First of all, one would like to have a measure for the centre of the distribution. The three measures used for this purpose are the MEAN, the MEDIAN and the MODE.

1. Mean

The MEAN (or arithmetic mean) is also known as the AVERAGE. It is calculated by totalling the results of all the observations and dividing by the total number of observations. Note that the mean can only be calculated for numerical data.

Example 6:

Measurement of the heights of 7 girls gave the following results:

141, 141, 143, 144, 145, 146, 155 cm (a total of 1015 cm for 7 measurements)

The mean is thus 1015/7, which is 145 cm.

2. Median

The MEDIAN is the value that divides a distribution into two equal halves.

The median is useful when some measurements are much bigger or much smaller than the rest. The mean of such data will be biased toward these extreme values. Thus the mean is not a good measure of the centre of the distribution in this case. The median is not influenced by extreme values. The median value, also called the central or halfway value, is obtained in the following way:

- List the observations in order of magnitude (from the lowest to the highest value or vice versa).
- Count the number of observations (n).
- The median value is the value belonging to observations number (n + 1) / 2 if n is odd or the average of the middle two numbers.

Example 8:

The weights of 7 pregnant women are 40, 41, 42, 43, 44, 47, 72 kg.

The median value is the value belonging to observation number (7 + 1)/2, which is the fourth one: 43 kg.

Note that the mean weight of this set of observations is 47 kg. This is an illustration of how the mean is affected by extreme values (in this case 72 kg) while the median is not. If the largest weight in this set of observations had been 51 kg instead of 72 kg, the median would still have been 43 kg, but the mean weight would have been 44 kg.

Note also that if there would be 8 observations: 40, 41, 42, 43, 44, 47, 49 and 72, the median would be 43.5 kg (the average of 43 and 44); the mean in this case would be 47.25 kg.

3. Mode

The MODE is the most frequently occurring value in a set of observations.

The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped.

In **Example 4** (number of patients treated for malaria at clinics) the mode is '0 to 19', as this outcome is recorded most frequently (25 times out of 80).

The mode can also be used for categorical data, whether they are nominal or ordinal.

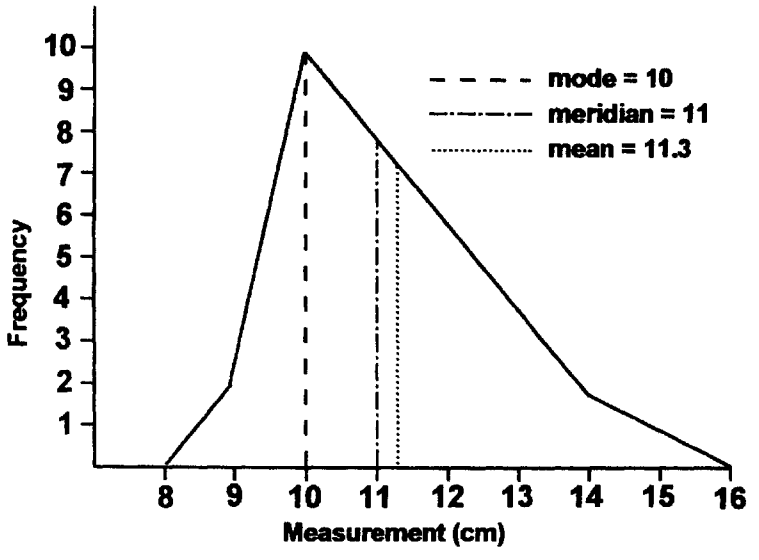
In **Example 1** (method of family planning) the mode is 'none'. In **Example 2** (number of clinics experiencing drug shortage) the mode is 'rarely'.

In summary, the mean, the median and the mode are all measures of central tendency. The mean is most widely used. It contains more information because the value of each observation is taken into account in its calculation.

However, the mean is strongly affected by values far from the centre of the distribution, while the median and the mode are not. The calculation of the mean forms the beginning of more complex statistical procedures to describe and analyse data.

Figure 22.6 shows a distribution curve in which the mean, the median and the mode have different values.

Figure 22.6: Mean, median and mode in a distribution curve.



GROUP WORK

1. Describe your sample(s) in terms of background variables (sex, age, etc). and dependent variables (e.g., defaulter/complier, user/non-user).

2. Make sure that you have made frequency counts for all variables in your study (from your data master sheets). Calculate percentages in relation to the total number of study units (or calculate proportions/ratios/rates where appropriate).

3. Check your objectives to determine which variables require frequency tables that should be included in your report. Usually frequency tables are presented for some of the background variables, the dependent variable(s) and the most important independent variables. Prepare the frequency tables.

4. Make histograms, bar charts, pie charts and/or line graphs, if useful. Prepare brief descriptions that interpret what each of the figures means.

5. Calculate means, medians and modes, if applicable, and interpret the results.

6. Familiarise yourself with the results and try to understand as fully as possible what they mean.

REFERENCES:

All epidemiology and statistics textbooks mentioned in **Modules 9** and **28**.

Trainer's Notes

Module 22: DESCRIPTION OF VARIABLES

Timing and teaching methods

1 hour	Introduction and discussion
3 hours+	Group work

Introduction and discussion

- It is likely that the participants will be familiar with some of the concepts introduced in this module, such as percentages and proportions. Moreover, at this stage, the groups will have already prepared frequency distributions (including calculation of percentages). Therefore these concepts should be only briefly mentioned in the presentation, especially if the knowledge level of the participants is high, so as not to lose their interest. However, special attention should be given to what to do with missing values when calculating percentages.

- Although definitions of percentage, proportion, ratio and rate are given in the module, it is more important to provide examples or ask participants to provide their own.

- When presenting **Example 3**, you might also ask participants to describe how they have grouped numerical data and discuss whether there were too few or too many categories.

- Examples should not be merely presented; they should be used in informal exercises. For example, ask participants what is the mean, median and mode of a given set of measurements, instead of providing them with the answers.

Group work

- Before the group makes frequency counts for all variables from the master sheets, have them review whether the data have been categorised correctly. Also be sure that the total number of informants (study units) for each group that has been studied has been defined.

- Remind the group that fully developed frequency tables are only for those variables that have to be described in the final report. Usually tables are needed for some of the background variables of the target group(s), and sometimes for the most important independent variables. Many of the other background and independent variables will be presented using cross-tabulations (**Module 24**).

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 23

ANALYSIS OF QUALITATIVE DATA

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 23: ANALYSIS OF QUALITATIVE DATA

OBJECTIVES

At the end of this session you should be able to:

1. **Describe** efficient ways of ordering and summarising qualitative data.
2. **Indicate** why it is essential to start summarising and analysing during the field work.
3. **List** the major steps in analysing qualitative data and drawing conclusions.
4. **Make an outline** of how you will proceed with the ordering and summarising of your qualitative data, and with the subsequent analysis.
5. **Plan** on how to report your qualitative data, integrated in the most effective way with your other data.
6. **Indicate**, either now or at the end of data analysis, what additional activities you will undertake to test or confirm your findings in order to prove their validity.

I. Introduction

II. Procedures for processing and displaying qualitative data

III. Drawing and verifying conclusions, using different data sets

IV. Reporting the data

V. Further strategies for testing or confirming findings to prove validity (optional)

I. INTRODUCTION

In previous **Modules (9, 10, 13)** it was pointed out that we use qualitative research techniques if we wish to obtain insight into certain situations or problems concerning which we have little knowledge. Qualitative techniques such as the use of **loosely structured interviews with open-ended questions, (focus) group discussions, observations, projective and participatory approaches** will therefore be appropriate in many studies, especially at the onset. For sensitive topics they may be the only reliable techniques.

Irrespective of how and for what purpose the data has been collected, the researcher usually ends up with a substantial number of pages of **written text** that needs to be analysed.

Although procedures and outcomes of qualitative data analysis differ from those of quantitative data analysis, the principles are not so different. In both cases the researcher will have to:

- describe the sample populations;
- order and reduce/code the data (data processing);
- display summaries of data in such a way that interpretation becomes easy, e.g., by preparing compilation sheets, flowcharts, diagrams or matrices;
- draw conclusions, relate these to the other data sets of the study and decide how to integrate the data in the report; and
- if required, develop strategies for further testing or confirming the (qualitative) data in order to prove their validity.

We will now examine each of these points in more detail.

II. PROCEDURES FOR PROCESSING AND DISPLAYING OF QUALITATIVE DATA

1. Description of the sample population in relation to sampling procedures

A useful first step in data processing (as well as in the reporting of findings) is a description of the informants. If numbers allow, relevant background data may be tabulated, for example on age, sex, occupation, education or marital status, as is the practice in quantitative studies.

However, as qualitative data originates from small samples (sometimes a handful of key informants or focus group discussions and observations) more information is required to place the data in its context.

> **For example**, who were the key informants, what made you decide to choose them? Who took part in the focus group discussions? How were the participants of the groups selected and how representative are they for your study population? For observations: under what circumstances were they carried out? Who were observed, and by whom?

Unless this type of information is provided, interpretation of data may appear haphazard.

2. Ordering and coding of data

We will discuss two types of qualitative data:

- answers to open questions, and
- more elaborate narratives from loosely structured interviews or FGDs.

(1) Answers to open questions

The most commonly collected qualitative data are the answers to open questions. They form part of every HSR study. When developing your protocol, in **Module 13**, you already did an exercise in the systematic ordering of such data: on the answers to the question 'Why are you smoking?' which we will discuss in depth again to analyse the different steps*.

(1) A first, basic step in the analysis of answers to open questions is to **list** the answers of a *sample of 20-25 informants* as they were provided (adding the questionnaire number in order to avoid losing the connection with the informant's other data).

(2) Then **read** the answers carefully, remembering the purpose of the question. The question 'why are you smoking' was supposed to help nursing students to develop an intervention against smoking.

(3) **Make rough categories** of answers that seem to belong together and **code** them with a key word. For example, answer 3 (It gives me pleasure) and answer 14 (I like to blow smoke rings) could be labelled with the term 'pleasure', which could be abbreviated with the code *pleas.*

(4) Then **list** again all answers but now **per code**, so that you get some 5-7 short lists, for example:

Pleasure (pleas)	*Being sociable (soc)*	*Giving self-confidence (selfc)*
2. I like the feel of the cigarette in my hand	10. All my friends are smokers	6. Because I feel confident and in-charge when smoking
3. Because it gives me pleasure	11. It helps to make people more friendly and comfortable, when offering a cigarette	7. It helps me to think better
5. I like to blow the smoke through my mouth and nose		18. It helps me to reduce the pressure and tension at work
14. I like to blow smoke rings		17. It helps me to relax
15. I like the taste		

(5) Then **interpret** each list, and end up with some 5-7 meaningful categories with a characteristic key word. For example: *Pleasure, being sociable, giving status, giving self-confidence, addiction, defiance.* There may be discussion on the need to split up some categories or combine others with few answers. Answers 17 and 18, for example could be put in a separate category *reducing stress.* In that case there would be seven categories. The category *defiance* may have two answers: 4. I do not see why I would give up smoking!! and 12. Why not?!! The exclamation marks indicate that defiance rather than lack of knowledge forms the motivation for the answer. Without this addition by the interviewer, these answers would have been difficult to code.

Now you can make a tentative interpretation according to the assumed willingness of your informants to change their behaviour. For those who smoke for *pleasure* or to *socialise* it might be most easy to give up smoking. Those who are *addicted* but tried to stop and those who feel they derive *status* from smoking might form a middle category, whereas for those who smoke to enhance their *self-confidence* and reduce *stress* or who are very *defiant* at the question why they smoke, it might be most difficult to stop.

* These steps have been adapted from Willms and Johnson (1996).

(6) Now try a next batch of 20-25 answers and **check if the labels work.** It is well possible that at this stage still some labels will be changed or that you decide to add new categories or combine others.

(7) **Make a final list** of labelled categories and code all data including the data you already processed with the abbreviated codes.

Then discuss whether you will stick to your tentative interpretation of the data and what this means for the content of the messages to address different reasons for smoking. This *content analysis* is the most important purpose of the analysis. By *counting* the answers under each label, however, the researcher will gain insight as well in how *common* the different reasons are.

(2) Elaborate narratives

The data from interviews with key informants or focus group discussions (FGDs) are as a rule more bulky than answers to open questions. The carefully transcribed field notes and tapes (see **Module 10.C** on FGD and **Module 13)** may consist of pages of narrative text. When analysing the texts we usually discover that, no matter how good our guidelines for the discussion were, the data contain valuable information but also a number of less essential details. In addition, the data is usually not presented in the order we need for our analysis, since informants may jump from one topic to the other.

To make the analysis easier, we have to **order** and **reduce** the data. Ordering is best done in relation to the objectives and the discussion topics. Again, it is best to systematically follow a number of steps.

(1) **Reread** your objectives and discussion topics

(2) Carefully **read** a number of the interviews, FGDs or narrative observations you want to process. Number the material according to the broad discussion topic it pertains to. Use a yellow marker to highlight particularly illustrative remarks. Use the margins to define sub-topics.

For example, in a gender and leprosy study carried out in different countries (used as example in **Modules 4, 8** and **11**) it appeared that the discussion topic *stigma* had to be differentiated according to different social settings in which it occurred: among close relatives (parents-children), spouses, in-laws, and community members. Further, a distinction had to be made between self-stigmatisation (e.g., a wife diagnosed as a leprosy patient encouraging her husband to marry a second wife in order to prevent divorce, or a patient not attending community meetings for fear of being avoided) and stigmatisation by others. Different degrees of severity in stigmatisation could also be distinguished, varying from slight avoidance to complete expulsion. If stigma would be topic (11) in your discussion list, you would mark everything related to stigma with an (11) in the margin, and add *key words* such as *self-stigm., spouse, in-laws, comm.,* in the margin, as well as key words such as *sleep(*ing) *sep(*arately) or *divorce* indicating the severity of the stigma. (See **Annex 10C.2** in Module 10C for an example.)

(3) **List all key words that belong to a certain topic in the sub-categories** that have been developed under (2). E.g., everything belonging to stigma could be subdivided and listed in the four major social settings in which stigma was found to manifest itself.

(4) **Interpret** the data, e.g., distinguish the major forms in which stigma manifests itself in these different social settings, try to make a ranking order of severity and link it to other variables (such as degree of deformity, socio-economic status) in order to understand differences in stigma.

(5) Then **code all your qualitative data** in this way. If necessary, adapt your coding scheme as you order, code and interpret more data. In that case, you should again read and possibly re-code the material you have already processed.

Note:

You may already have analysed and coded your qualitative data in the field (as advised in **Module 13**), in order to adjust and deepen your interview guides or topic lists. In that case it may be possible to develop your final coding list in one cycle instead of two.

However, instead of developing a very detailed coding system on your rough data, you may also refine your interpretation as you record your roughly coded, summarised data in **COMPILATION SHEETS.**

3. Summarising data in compilation sheets

After ordering the data we will have to summarise them. A useful first step is summarising all data of each study unit per study population on separate compilation sheets.

Like the master sheets for quantitative data, compilation sheets for qualitative data consist of a number of columns with the topics covered by the study as headings. These may be further sub-divided in smaller themes that you identified and coded when ordering the data (see **Annex 23.1**). Each interview, FGD or observation gets a number and is successively entered in that sequence on the relevant compilation sheet. If there are different categories of informants within one study population, for example, young mothers and an older generation of mothers, or male and female patients, the data for these groups are entered on separate sheets. If the topics covered in those sub-groups are not completely identical, it is important to be systematic and follow roughly the same sequence of topics for each category of informants. The information inserted is summarised in key words and key sentences, clear enough to remember the statements informants made. (As the number of each study unit is entered in the compilation sheet, it is always possible to go back to the original data and present the full statement, for example in a presentation or in the research report).

Now you have an overview of all data per study population on one or more big sheet(s). If you read the columns, you have a list of answers of all group members on a certain (sub-)topic. If you read horizontally, you can per informant relate different topics to each other or to personal characteristics of the informant. It becomes also easy to compare the answers of different groups on specific issues by comparing compilation sheets.

> **For example,** in **Annex 23.1,** the personal data of leprosy patients (recently declared cured) and a number of topics and sub-topics discussed with them are presented. Stigma actually experienced, which originally was one topic, has in the compilation sheet been subdivided in the four major social settings in which stigmatisation may occur: close blood relatives, marriage, wider circle of spouse's relatives and community. In each of those still finer distinctions can be made (e.g., community can be neighbours, friends, work mates, school mates or distant community members). As samples are small, these may all be inserted under the heading 'community'. Codes *(italics)* can be added to the statements presented in key words, for example *big fear* and *worried* under the heading 'first reaction'. From the three examples presented, it already appears (confirmed by the analysis of all data in all four countries) that in general the stigma feared when patients hear the diagnosis of leprosy is bigger than the stigma in reality experienced. Patient (12) is in this respect an exception. Ironically, the husband who divorced her had already died from another disease at the moment she was declared cured from leprosy. Horizontal comparison of the data of patient (1) teaches us that it is highly unlikely that the man's friends do not know about the disease, as even after he has been declared cured he has visible signs. Here the researchers had to interview the friends to find out if indeed this man was (or had not been) stigmatised at all by the community.

You may notice that interpretation of data and labelling becomes indeed easy when using compilation sheets, as a researcher can visualise all aspects of his/her informants even if (s)he looks at one aspect at a time for the whole study population.

A next step in summarising may be the combination, contrasting or further analysis of important topics through graphical displays such as matrices, diagrams, flow charts and tables.

4. Further summarising of data in matrices, figures and tables

Matrices

Matrices can be used for quantitative as well as qualitative data comparison. In qualitative data we may compare different groups or data sets on important variables, presented in key words.

> A MATRIX is a chart that looks like a cross-table, but contains words (as well as, sometimes, numbers).

In a focus group discussion on changing weaning practices, the researchers listed the answers of young mothers concerning the introduction of soft foods and those of mothers above childbearing age. They then summarised these answers in a matrix:

Figure 23.1: Matrix on introduction of soft baby foods among mothers of different age groups

AGE GROUPS	ONSET SOFT FOOD	TYPE OF FOOD	FREQUENCY OF SF/DAY
Young mothers (20-30 years)	Range: 4-7 months Average: 6 months	• Soft porridge • Soft porridge with pounded groundnuts • Mashed potatoes, mashed fruits, soaked biscuits	2-4 times daily • Depends on availability of mother and caretaker • Depends on appetite of child
Mothers past child-bearing age (>45)	Range: 5-11 months Average: 8.5 months	• Soft porridge • Soft fruit	1-2 times daily • Depends on availability of mother and caretaker • Depends on appetite of child

This type of display made it easy for the researchers to conclude that:

- younger mothers start giving soft foods, on average, 2.5 months earlier than the generation of their own mothers;
- younger mothers use a larger variety of soft weaning foods than women in the preceding generations; and
- younger mothers give soft foods to their babies more frequently, but for the same reasons as their mothers did.

Matrices facilitate data analysis considerably. They are the most common form of graphic display of qualitative data. They can be used to order and compare information in many ways, for example, according to:

- time sequence (of procedures being investigated in different periods, for example),
- type of informants (as in the example above), or
- location of data collection (to visualise differences between rural and urban populations).

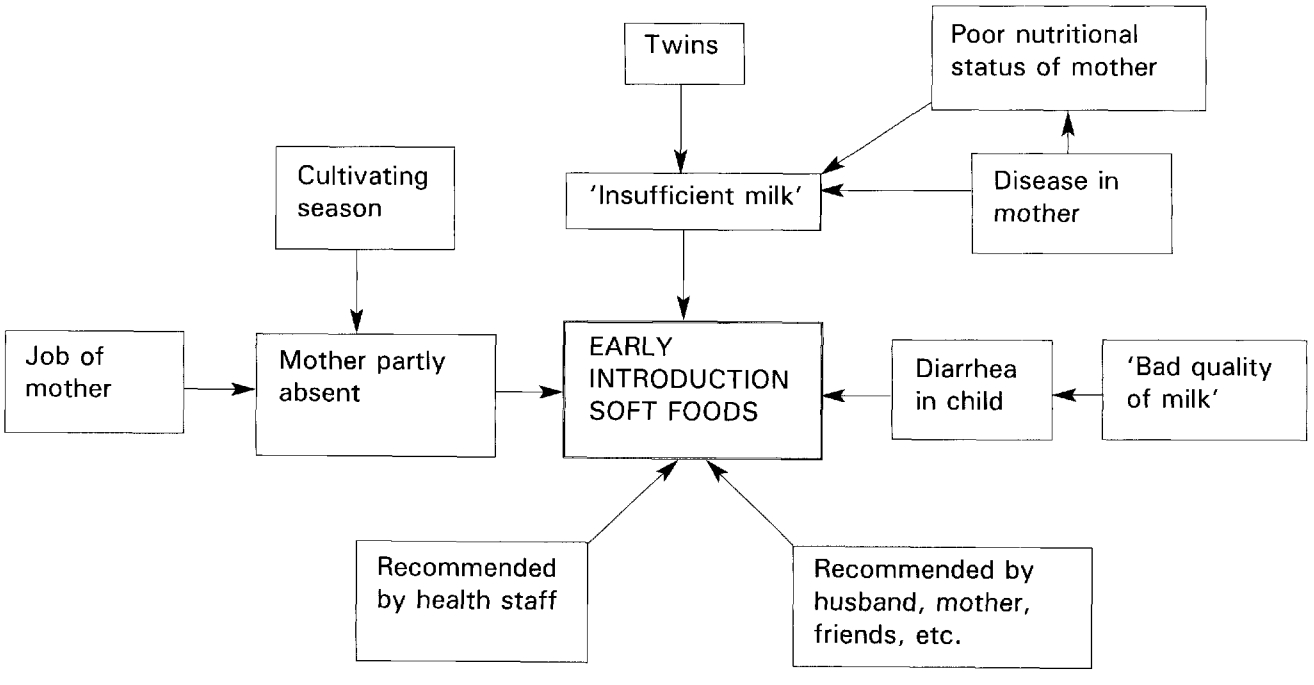
Diagrams

A DIAGRAM is a figure with boxes containing variables and arrows indicating the relationships between these variables.

When analysing the problems you wanted to investigate during the development of your protocols, most groups developed a diagram. In a similar way diagrams can be developed to summarise findings of a study. (See **Figures 23.2** and **23.3**).

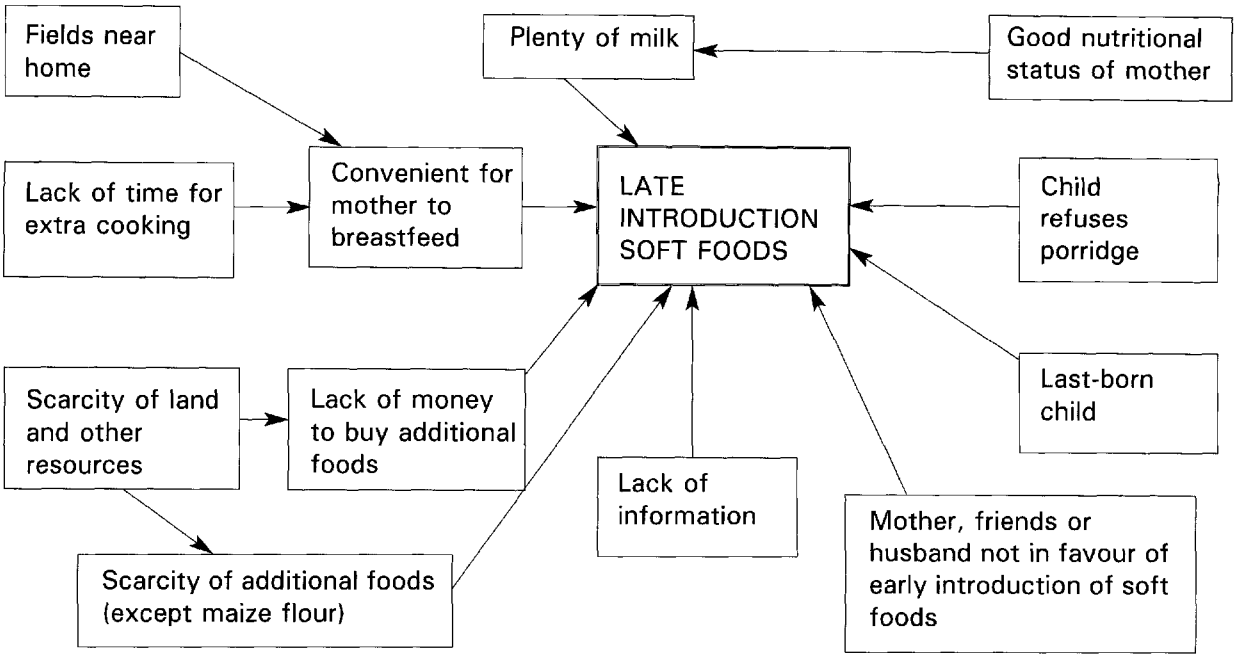
You might use a diagram to illustrate a crucial issue in your study, combining all available qualitative and quantitative data collected.

Figure 23.2: Reasons for early introduction of soft foods by young mothers



Diagrams, like matrices, can be of great assistance in providing an overview of the data collected and in guiding data analysis.

Figure 23.3: Reasons for late introduction of soft foods by young mothers



Flow charts

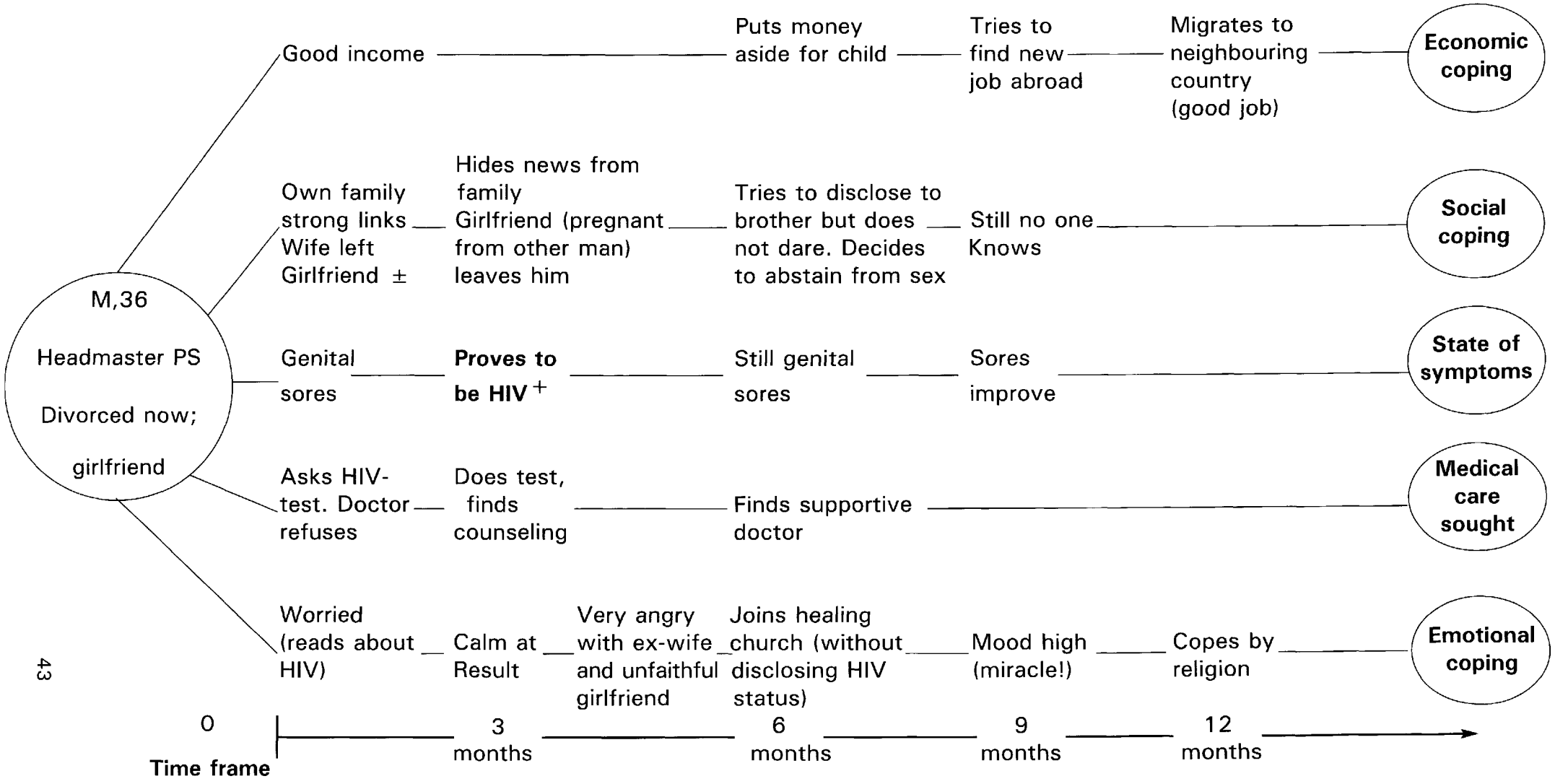
FLOW CHARTS are special types of diagrams that express the logical sequence of actions or decisions.

The figure preceding **Modules 1-18,** indicating the successive steps in protocol development, is an example of a flow chart.

Flow charts are especially useful to summarise different flows of events that are mutually connected. A counselling team in Bulawayo, Zimbabwe, for example, which interviewed some 95 HIV positive persons in-depth over a period of two years, summarised the roughly 100 pages of interview material for each informant by drawing five lines (see **Figure 23.4**). One central line presented the development of the disease over time, with crises and periods of relative well-being. Another line presented different forms of medical care sought, a third the flaws in economic status connected to the disease (e.g., loss of job, seeking employment elsewhere), a fourth the possible changes in social status such as divorce or (re)marriage, whereas a fifth line presented the patient's emotional status linked to events occurring in the four other fields (e.g., positive coping, depression). These flow charts were extremely useful for comparison of data, per informant and between different groups of informants (e.g. males/females, single/married). They highlighted the impact of the disease on the lives of different groups of patients and their way of coping with it.*

* Meursing (1997) *A world of silence.*

Figure 23.4: Flowchart on coping of HIV$^+$ persons with their condition over time



Example of relatively well-to-do man who copes in solitude despite supportive relatives because he is too ashamed to unclose his HIV$^+$ status.
Adapted from Meursing K (1997) *A world of silence; Living with HIV in Matabeleland, Zimbabwe.* Amsterdam: Royal Tropical Institute.

Tables

A TABLE is a chart with rows and columns that has numbers in the various cells or boxes.

Qualitative data can also be categorised, coded, inserted in master sheets or computer and *counted*, together with other quantitative data, and displayed in tables. Answers to open-ended questions in questionnaires will usually be categorised and summarised in this way. However, you will in the first place want to *analyse the content* of the individual answers in each category. (See section II-2 and section III in this module.)

III. DRAWING AND VERIFYING CONCLUSIONS

Drawing and verifying conclusions is the essence of data analysis. It is not an isolated activity, however. When we start summarising our data in compilation sheets, flowcharts, matrices or diagrams, we continuously draw conclusions, and modify or reject quite a number of them as we proceed. Writing helps generate new ideas as well. Therefore **writing should start as early as possible**, right from the onset of data processing and analysis, if only for ourselves. No creative insights should get lost!

Note:

Collection, processing, analysis and reporting of qualitative data are closely intertwined, and not (as is the case with quantitative data) distinct successive steps. It may often be necessary to go back to the original field notes and verify conclusions, collect additional data if available data appear controversial, and get feedback from all parties concerned.

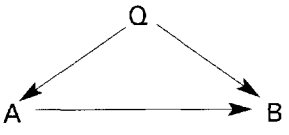
Identifying variables and associations between variables

In **Module 8** we stated that sometimes we do not know enough about a situation to define variables beforehand. Only during or at the end of the study it will be possible to define certain variables and search for associations with other variables, without having the prior aim of *measuring* them. Many HSR studies have qualitative parts with open questions, key informant interviews, focus group discussions or observations for the purpose of identifying these variables. The researcher who uses such a qualitative approach should be like a detective who searches for evidence, accounts for countervailing evidence, and verifies the findings by looking for independent, supporting evidence, until (s)he is confident about possible associations among certain variables which shed light on the problem under investigation.

For example, if we find among the mothers who wean their children early that quite a number have jobs, we may assume that having a job contributes to early weaning. Similar studies carried out elsewhere with similar findings support this assumption (independent evidence). Only if there are very few employed women who wean their children late, however, can we be more certain that our assumption is true, and for each of those exceptions we should try to find an explanation. Do the mothers take their children with them (crèche at place of work) or do they work near their homes so that they can feed the baby during breaks? Or do they successfully combine breast-milk with alternatives? If yes, why don't more mothers try this combination? etc. etc.

Finding confounding or intervening variables

Sometimes variables appear to be related but the association cannot easily be explained. Other times it seems that variables should logically go together, but you cannot find a relationship. In cases such as these there may be another variable ('Q') influencing the association between the two variables concerned, that has to be identified (see **Modules 8, 9** and **26**).



For example, one expects a relationship between the quality of drinking water and the incidence of diarrhea. It is assumed that the incidence of diarrhoea would decrease as the number of water faucets in a village increased. If there is no change over time, there might be a confounding variable. People, for example, may dislike the taste of tap-water so much that they use it for everything, except for drinking.

Note:

Such unexplained associations may appear in any study. The essential characteristic of a qualitative research approach is that it purposively looks for such associations **during** the fieldwork, and that additional questions and tools may be developed to highlight such relationships. In quantitative surveys that attempt to objectively measure the strength of a *presupposed* association between two variables, the tools should not be changed once the fieldwork is ongoing.

Integrating qualitative and quantitative data

Thus far we have discussed the analysis of qualitative data as a separate activity. However, if a research team has collected qualitative as well as quantitative data, which is the case in most HSR studies, it would be foolish not to look at them in combination, as this can inspire to deeper and more rewarding analysis.

For example, the Indonesian 'gender and leprosy' research team found, when analysing the registration data of 4500 new leprosy patients who had registered over the past five years, that the M/F ratio was most unfavourable in the age group of 15-44 years. This was a puzzling finding, as in Nepal women in this age group were reporting much better (though still less than men). In-depth interviews with staff revealed that they suspected adolescent girls and young women to hide their skin patches, because of shameful associations with dirt, ugliness. This provided the incentive for a further break down of the quantitative data, which revealed that the M/F difference in reporting was indeed most pronounced in the 15-34 age group, and levelled off above 35. The reason(s) for this relatively large gender difference in the younger age groups were then further explored.

Content analysis of qualitative data for action

Quantitative data serve in the first place to convince health authorities **that** there is indeed a serious, sizeable problem; qualitative data help to provide ideas on **how** to solve it. The FGDs on weaning foods with young mothers and mothers who had surpassed the childbearing age, for example, will yield many suggestions on how to develop interventions with the mothers which they are likely to consider useful and will be able to implement. Likewise, the in-depth interviews with leprosy and ex-leprosy patients will provide new insights into how best to counsel new patients and their close relatives/spouses in order to reduce unnecessary fears.

Computer analysis of qualitative data

With the ever-increasing importance of computers in research, strategies for analysing qualitative data by computer have been/are being developed. There are several possibilities, ranging from simple word processing programs to highly sophisticated Qualitative Data Management Software including possibilities for statistical testing of associations. As numbers are usually small in HSR and content analysis, which can be done by hand, is most likely more important than *testing* of associations, we will not elaborate these techniques here. Rather we refer the interested students to Anthropology or Psychology Departments at universities that have experience with programs such as Qualitan or SPSS for qualitative data processing.

IV. REPORTING THE DATA

Basically, there are two ways of reporting qualitative data that form part of a study in which different research techniques were used. One way is summarising the major qualitative results in a separate section of the findings, with examples and quotations, following the objectives that guided the collection of this particular data. The results would then be discussed in the chapter 'Discussion', together with the results of other, more quantitative data collection tools and would subsequently be reflected in the summary of the findings and the recommendations.

Another possibility is to fully integrate different data sets in the chapter of findings, ordered according to the objectives of the entire study. If quantitative and qualitative data have been analysed and sometimes even collected in an integrated way, it would also be logical to present them in an integrated fashion. Attention should be paid that no valuable data get lost. Therefore a rough draft of all important findings is required in any case, after which can be decided to present the data either in separate sections or chopped up for integration with other data. (For details see **Module 32.**)

V. FURTHER STRATEGIES FOR TESTING OR CONFIRMING QUALITATIVE FINDINGS TO PROVE VALIDITY

Researchers who use quantitative research designs reduce their data to numbers and apply statistical tests. This does not necessarily insure that their research results are valid: something may have gone wrong during sampling or collection of data or even in the earlier design of the study (overlooking possible confounding variables). The following strategies will therefore be of use to any researcher. They are particularly relevant, however, to qualitative research, since the small numbers of qualitative data often generate questions concerning its validity.

1. **Check for representativeness of data.**

   Although in qualitative research informants have usually not been selected randomly, they must have been selected systematically, according to previously established rules. (See **Module 11.**) Check whether you have indeed interviewed all categories of informants needed to get a complete picture of your topic (not relying excessively on talkative authorities). Make sure that you do not generalise from unrepresentative events.

2. **Check for bias** due to observer bias or the influence of the researcher on the research situation. We discussed this in detail in **Module 10**.

3. **Cross-check data with evidence from other, independent sources.**

   These sources may be different independent informants, different research techniques employed to investigate the same topic, or results from other, similar studies. (See **Modules 5, 9** and **10.**) The data should confirm or at least not contradict each other.

Actively cross-checking data, looking for independent evidence or counter-evidence, is one of the most important ways to enhance the validity of research data.

> **For example**, answers of husbands and wives (and other informants concerned) should confirm each other on such issues as who decides whether and what family planning methods should be used, who decides whether daughters should be circumcised, or what has changed in husband-wife relationships after the diagnosis of leprosy or another feared disease in one of the spouses.

4. **Compare and contrast data.**

   Comparison will often have been built into the research design through including different categories of informants.

   > If we want to be sure, **for example**, that variable A (high level of education) influences variable B (use of family planning methods) we have to compare a group of mothers with high education to a group of mothers with low education on their use of family planning methods.

   Comparing and contrasting data is important if you are attempting to **identify** your variables as well as to **confirm** associations among variables.

5. **Use extreme (groups of) informants to the maximum.**

   In the discussion of study design and sampling we stated that it may be useful to look for categories of informants that represent the extremes on a certain variable.

   > **For example**, you may find it most useful to study 'drop-outs' and regular attendees of TB services, leaving out the category of irregular attendees. This may be the most efficient way of identifying the key variables that influence the attendance behaviour of TB patients.

6. **Do additional research to test the findings of your study.**

   The results of your study may be so intriguing that you decide to do a follow-up study afterwards. Such a study may be undertaken for several reasons:

   - to replicate certain findings,
   - to rule out (or identify) possible intervening variables,
   - to rule out rival explanations by investigating them, or
   - to look for negative evidence.

   Additional studies undertaken for one or more of these reasons may serve to make the results of your original study more convincing.

7. **Get feedback from your informants.**

   Throughout **Modules 1 - 20** we have stressed that you need to involve all parties concerned in the various stages of the research. This is important not only for ethical reasons or because it will improve the chances that the results will be implemented, but also because **it will improve the quality of your study design, of your data, and of the conclusions drawn from these data.** Suggestions and additional information collected during feedback sessions will invariably increase the quality of your research report.

GROUP WORK (Time needed will depend on the amount of qualitative data collected.)

1. **Check whether you listed all sources of qualitative data** for each objective when, in the group work session of **Module 21**, you made an inventory of all your data.
2. **Ensure** you have finished the **categorisation of answers to all your open questions**, included them in your master sheets or computer analysis together with other data sets, and that you then do a content analyses of the answers for inclusion of relevant data in recommendations for action or subsequent action plans (see **Module 33**)
3. If your study included FGDs, interviews with key informants or observations: **describe your samples.**
4. Organise this data by topic, further code it, if necessary, **and enter the data by topic on compilation sheets.**
5. **Decide** whether you will use matrices, diagrams and/or flowcharts to summarise your data.
6. **Interpret the data**, comparing different groups of FGD participants or key informants (if you have them) and see how they answer your research objectives.
7. **List the major findings and conclusions of the qualitative data** and determine how they complement data from other sources in your study.
8. **Decide on how you want to enter the data in your report:** either in one section or integrated with the findings collected through other data-collection techniques. Decide what should come in the discussion and what is useful material for developing interventions.
9. **Verify your conclusions** (see section V) and decide whether and how you would like to further test certain conclusions.

REFERENCES:

Miles MB and Huberman AM (1984) *Qualitative data analysis, a sourcebook of new methods.* Beverley Hills, CA, USA.: Sage Publications.

Patton MQ (1990) *Qualitative Evaluation and Research Methods*. 2nd ed. Newbury Park, CA: Sage Publications.

Spradly JP (1979) *The ethnographic interview*. New York, NY, USA.: Holt, Rinehart and Winston.

Walker R (ed) (1985) *Applied qualitative research*. Hants, UK: Gower Publishing Company Ltd.

Willms DG and Johnson NA (1996) *Essentials in Qualitative Research: A Notebook for the Field*. Hamilton, Canada: Mc Master University.

Yin RK (1984) *Case study research: design and methods*. Beverly Hills, CA, USA.: Sage Publications.

NB: A major source of inspiration for writing this module was Miles and Huberman's book. Section V of this module is a heavily abbreviated and adapted version of their chapter VII.

Annex 23.1: Example of compilation sheet (gender and leprosy)

Nr	Personal data					Symptoms		First reaction(10)	Stigma in reality experienced (11)				Ec./domestic act (12)	Perception of cure (25)
	Sex	Age	Educ.	Marr.	Ec. status.	at diagn. (6)	now (18)		Spouse	Relatives	In-laws	Comm.		
1	M	40	6yrs	Yes	Farmer Shopkeeper	• Patches • Painful nerves • Dropfoot *(2)*	↓ *None* *Still*	*Big fear* • Wife will run away • Community will isolate him • Fingers and toes will drop off • No longer able to work and sustain family	• Remains *supportive* • Helps more in shop • He decided to abstain from sex (8 months) *Self-stigm.*	• Children *supportive;* small ones not aware • Parents & Br/Si visit + mix as before	Not told, *hiding*	• Not told • Thinks friends don't know • Behave as before *Hiding.* *No stigma?*	Hires labour (No force to farm) *Income* ↓	*No* (still signs: drop-foot)
10	F	21	8yrs	Yes	Hu. farms Fa. big farmer	Patches (teacher saw and referred her to HC)	*None*	*Knew little; worried* • Bad disease • Fiancé will break off marriage proceedings	• Fiancé inquired at HC: - if curable - if she could get children • Marriage postponed till patches subsided • She now has child *Stigma reversed*	• Parents very *supportive*	First wife of hu. told in whole village	All villagers came to marriage. *No stigma*	Does everything	Yes (No signs)
12	F	60	-	Yes, but now *divorced*	Small farming + trade	Patches	↓	*Worried* • Bad disease • Hu. angry	Hu. kicked her out. *Divorce*	Son took her in. *Supportive*	Da.-in-law *supportive*	Avoided big meetings but now OK *Self-stigma reversed*	Small trade to earn bus fee for treatment	Yes, cured (but hu. died!)

Trainer's Notes

Module 23: ANALYSIS OF QUALITATIVE DATA

Time and teaching methods

1 hour Introduction and discussion
1 hour Group work

(Time should be adjusted, depending on amount and type of qualitative data.)

Introduction and discussion

- If none of the groups have qualitative data other than some open-ended questions in questionnaires, you may wish to concentrate on sections II, to give participants an overview of how one could process qualitative data, and only briefly touch on sections III and IV.
- However, if the course participants have experience/training/interest in research, you might fully cover sections III and IV, even if they have not collected large quantities of qualitative data. The procedures presented for drawing conclusions and testing validity are pertinent to all types of research, and the methods for checking and cross-checking data may not be known to all participants.
- Refer to the analysis diagram each group made when preparing its research proposal, to the flow chart in front of **Modules 1-18** and **22-33**, and present any other examples of charts or graphs you find illustrative on overhead sheets or flip charts.
- If one or more groups have done extensive qualitative research, cover the module in detail, using examples from their studies. Most likely none of the participants will be very familiar with analysis of qualitative data.
- Let the groups that have done qualitative research describe in plenary how they analysed the data from focus group discussions, observations, and/or interviews with key informants. Ask what additional questions they added or questions they dropped in the course of successive interviews, and why.

Group work

- **For all groups:**

  Check whether any of their open-ended questions require analysis of the content of individual answers. Some opinion questions might provide valuable illustrative material for their reports. Take note of this, as the groups might forget such data when they get involved in tables and statistics. Discuss whether merely listing the statements is sufficient for content analysis or whether graphic display of the data would be desirable.

For those groups that have elaborate qualitative data from FGD or key informant interviews:
Review all the data available with them and assist them in following the group work directions.

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 24

CROSS-TABULATION OF QUANTITATIVE DATA

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 24: CROSS-TABULATION OF QUANTITATIVE DATA

OBJECTIVES

At the end of this session you should be able to:

1. **Describe** the difference between descriptive and analytical cross-tabulations.
2. **Construct** all the important cross-tabulations which will help meet your research objectives.
3. **Interpret** the cross-tabulations in relation to your objectives and study questions.

I. INTRODUCTION

Thus far we have made tables containing frequency distributions for **one variable at a time** in order to partially describe our data. Depending on the **objectives** of our study and the **study type**, we may have to **examine the relationship between several of our variables** at the same time, in order to adequately describe our problem or identify possible explanations for it.

For this purpose it is appropriate to construct CROSS-TABULATIONS.

II. DIFFERENT TYPES OF CROSS-TABULATIONS

Depending on the objectives and the type of study, different kinds of cross-tabulations may be required:

- There are cross-tabulations that aim at **describing the problem under study** by presenting a combination of variables. Exploratory and descriptive studies will generate such tables, but also in analytic studies it is necessary to first describe the problem under study (see **Tables 24.2** and **24.3**).

- Other cross-tabulations **analyse the relationships** between independent and dependent variables, i.e., **between defined problems and factors contributing to those problems**. Exploratory studies that identify contributing factors to problems may generate such analytical cross-tables, and cross-sectional comparative studies, case-control studies and cohort studies will automatically require them (see **Table 24.6** and **Tables 24.4** and **24.5** for the identification of associations).

- Then there are cross-tables, which aim at **comparing the outcome of interventions/ experiments** in a study group that participated in the intervention and a control group that did not participate, to determine whether the intervention made a difference. This is the standard procedure in experimental and quasi-experimental studies (see **Table 24.7**).

- Very common cross-tables, required in **any** type of study, are those that **describe the sample(s)** taken from the study populations on which the research concentrates. We will start with those.

1. Cross-tabulations to describe the sample

In any study, whether small scale (exploratory) or large scale, it is common to first describe the research subjects included in the sample(s) before presenting the actual results of the study. This can be done for separate variables in a simple frequency table (as shown in **Module 22**) or for a combination of variables in a cross-table.

Example 1:

A study was carried out on the degree of job satisfaction among doctors and nurses in rural and urban areas. To describe the sample a cross-tabulation was constructed which included the sex and the residence (rural or urban) of the doctors and nurses interviewed. This was useful because in the analysis the opinions of male and female staff had to be compared separately for rural and urban areas.

Table 24.1a: Type of health worker by residence

Residence	Type of Health Worker				Total	
	Doctors		Nurses			
Rural	10	(16%)	69	(38%)	79	(33%)
Urban	51	(84%)	113	(62%)	164	(67%)
Total	61	(100%)	182	(100%)	243	(100%)

Table 24.1a shows that a higher percentage of nurses than of doctors work in rural areas, but that, overall, a greater proportion of staff works in urban areas (67%).

Table 24.1b: Sex of health workers by residence

Residence	Sex of health workers				Total	
	Male		Female			
Rural	54	(43%)	25	(21%)	79	(33%)
Urban	71	(57%)	93	(79%)	164	(67%)
Total	125	(100%)	118	(100%)	243	(100%)

It can be concluded from **Table 24.1b** that there are more males serving in rural areas than females. These males in rural areas are apparently nurses.

To obtain an overview of the distribution of doctors and nurses by gender in rural and urban areas, we can construct the following two-by-four cross-table.

Table 24.1c: Residence and sex of doctors and nurses

Health workers		Residence				Total	
		Rural		Urban			
Doctors	Males	8	(10%)	35	(21%)	43	(18%)
	Females	2	(3%)	16	(10%)	18	(7%)
Nurses	Males	46	(58%)	36	(22%)	82	(34%)
	Females	23	(29%)	77	(47%)	100	(41%)
Total		79	(100%)	164	(100%)	243	(100%)

This table shows us at one glance that indeed male nurses dominate the rural health services. It also indicates that males dominate in the medical profession, (18% M <-> 7% F doctors) but that overall there are more female than male nurses, of whom the females mainly cluster in town.

The data in tables is usually listed in absolute figures as well as in relative frequencies (percentages or proportions).

As already seen in **Module 22**, for numerical data (such as age) the mean, median and/or mode may be calculated as well to describe the sample.

2. Descriptive tables to describe a problem

It is common to describe the problem on which a study concentrates. Cross-tables are very helpful in this respect.

Example 2:

We want to know the ages at which **teenage pregnancies** occur and whether they are more frequent among schoolgirls than among girls who are not attending school. In order to answer these questions we may construct the following cross-tabulation. (The data are imaginary.)

Table 24.2: Number of teenage pregnancies at different ages among girls attending school and not attending school (Province X, 1998 - 2000)

Age at onset of pregnancy	Number of Pregnancies		TOTAL
	Girls attending school (N = 500)	Girls not attending school (N = 500)	
12 years	2 (3%)	1 (1%)	3 (2%)
13 years	2 (3%)	3 (3%)	5 (3%)
14 years	5 (7%)	12 (12%)	17 (10%)
15 years	23 (34%)	35 (35%)	58 (10%)
16 years	36 (53%)	48 (49%)	84 (50%)
TOTAL	68 (100%)	99 (100%)	167 (100%)
Prevalence	68/500 = 13.6%	99/500 = 19.8%	167/1000 = 16.7%

Table 24.2 reveals that pregnancies occur already from 12 years onwards, but that from 15 years onwards the problem increases sharply in both girls who attend and who do not attend school. However, the percentage of pregnancies is higher among girls not going to school (almost 20%) than among girls who go to school (13,6%), and the first group appears to get pregnant at a slightly younger age.

Example 3:

A study was done to examine the factors contributing to the high proportion of stillbirths in a hospital. The following cross-tabulation describes how many of the fresh and macerated (wasted) stillbirths weighed less than 2500 grams and how many weighed 2500 grams or more.

Table 24.3: Weight of foetus by condition at birth

Weight of foetus	Condition at birth		TOTAL
	Fresh	Macerated	
Less than 2500 grams	9 (10%)	20 (63%)	29
2500 grams or more	79 (90%)	12 (37%)	91
Total	88 (100%)	32 (100%)	120

One can see from **Table 24.3** that most of the fresh stillbirths (90%) were of normal birth weight while a large proportion (63%) of macerated stillbirths had a low birth weight.

Note:

That in the descriptive cross-tabulations, it is convenient to put the groups that are to be described in columns (condition of birth for **Table 24.3**)

In **cross-sectional surveys**, it is possible to play around with variables and identify possible associations/relationships.

Example 4:

In a **cross-sectional survey on malnutrition**, for example, relationships could be tested between the duration of breastfeeding and the mothers' age, or the mothers' working status (answering previously formulated research questions, but sometimes new questions that crop up during analysis of the material).

Note that in such tables it is allowed to calculate your percentages both horizontally and vertically as all variables have a similar chance of appearing in the survey. However, we will usually put the variable that is assumed to influence the other one in rows, while the 'dependent' variable will be put in columns (see **Tables 24.4** and **24.5**)

Table 24.4: Duration of breastfeeding by mothers' age

Age (years)	Duration of breastfeeding			TOTAL
	0-5 months	6-11 months	≥ 12 months	
15-19	18 (62%)	8 (28%)	3 (11%)	29 (100%)
20-24	27 (44%)	25 (40%)	10 (16%)	62 (100%)
25-29	15 (18%)	33 (40%)	35 (42%)	83 (100%)
30-34	5 (14%)	13 (37%)	17 (49%)	35 (100%)
35-39	2 (10%)	7 (33%)	12 (57%)	21 (100%)
40+	0 (0%)	3 (30%)	7 (70%)	10 (100%)
Total	67 (28%)	89 (37%)	84 (35%)	240 (100%)

Table 24.4 suggests that the younger mothers breastfeed for a shorter period than the older mothers, as the highest percentage of mothers that breastfeed only 0-5 months is the youngest age group (15-19). With age the percentage of mothers who wean their children before they are 6 months decreases. For mothers who breastfeed more than 12 months we observe exactly the opposite trend: the highest percentage (70%) is in the oldest age group of mothers (above 40 years), whereas only 10% of mothers 15-19 years breastfeed longer than 12 months. **So there appears to be an association between age of mother and duration of breastfeeding.**

If you would like to determine whether there is an association between the working status of mothers and the duration of breastfeeding, **Table 24.5** would be appropriate.

Table 24.5: Working status of mothers in relation to duration of breastfeeding

Mothers' working status	Duration of breastfeeding			TOTAL
	0-5 months	6-11 months	12+ months	
Full time employed	56 (42%)	49 (38%)	27 (20%)	132 (100%)
Part time employed	5 (21%)	15 (62%)	4 (17%)	24 (100%)
Not employed	6 (7%)	25 (30%)	53 (63%)	84 (100%)
Total	67 (28%)	89 (37%)	84 (35%)	240 (100%)

The fully employed women breastfed for a shorter period than the not employed. Therefore, an association seems to exist between mothers' working status and the duration of breastfeeding.

Here we are in a grey field between descriptive and analytic tables, but we have seen in **Module 9** (study types) that cross-sectional surveys can easily turn into comparative (analytic) studies if sufficient data about possible contributing factors to the problem under study is collected (see below).

3. Analytic cross-tabulations

In **cross-sectional comparative** and **case-control studies** we compare two groups, one with a selected problem and one without, to identify independent variables that contribute to the problem.

You will remember from **Module 9** that cross-sectional comparison is based on a larger cross-sectional survey (which is a descriptive study). However, it can form the basis for a comparative study when we select extreme groups from the survey sample, one with a specific problem (such as severe malnutrition) and a control group (of well-nourished children) to identify contributing factors to the problem (of malnutrition).

Example 4 (continued)

One of the possible contributing factors to malnutrition of under 5's is knowledge of the mothers of appropriate weaning foods. The cross-sectional comparative study on malnutrition based on the survey gave the following results:

Table 24.6: Mothers' level of knowledge and nutritional status of their children

Level of mothers' nutritional knowledge (weaning foods)	Nutritional status of children				TOTAL
	Severely malnourished (cases)		Well-nourished (controls)		
Low	45	(69%)	10	(15%)	55
High	20	(31%)	55	(85%)	75
Total	65	(100%)	65	(100%)	130

It seems that the mothers of severely malnourished children have far less knowledge of weaning foods than the mothers of well-nourished children.

Other analytic cross-tabulations can be constructed for the study mentioned in **Example 4.** Each time, the two groups (severely malnourished children and well-nourished children) can best be systematically displayed in the columns. Different independent variables will then be put in rows, such as source of drinking water (protected or unprotected) or immunisation status (fully immunised or not).

Case-control studies follow the same procedures for the construction of cross-tabulations as outlined in **Table 24.6**.

In **cohort studies** cross-tabulations will be constructed in the same way as in case-control or cross-sectional comparative studies. The previously defined risk factors will be placed in the rows, whereas the participants in the study who develop certain conditions or diseases and those who remain unaffected will be placed in the columns of the table.

In **quasi-experimental** or **experimental studies**, a researcher will compare two similar groups, (one being subjected to an intervention and the other not) before and after the intervention in order to measure its effect.

If the nutritional cross-sectional comparative study mentioned in **Example 4** will lead to a health education intervention, it would be worthwhile to conduct a quasi-experimental study through which the effect of the health education effort can be measured (see **Table 24.7**).

Table 24.7: Women's attendance at nutritional education and their level of nutritional knowledge

Attendance at nutritional education	Level of nutritional knowledge			TOTAL
	Low	Average	High	
Attenders	20 (13%)	90 (60%)	40 (27%)	150 (100%)
Non-attenders	40 (45%)	30 (33%)	20 (22%)	90 (100%)
Total	60 (25%)	120 (50%)	60 (25%)	240 (100%)

Table 24.7 indicates that women who attended nutritional talks have a higher level of nutritional knowledge than non-attenders. It appears therefore that the intervention was effective.

Note that in the above cross-tabulation the groups that are to be compared have been put in rows, whereas the different levels of nutritional knowledge were put in columns. This is because nutritional knowledge is the outcome of the attendance at nutrition talks and it is therefore considered to be the dependent variable. The percentages are calculated horizontally as you are interested in comparing the level of knowledge between attenders and non-attenders.

To help you to interpret collected data and write your report in a systematic way, it is suggested that the dependent variables be placed in the columns while the independent variables are placed in the rows.

Note the difference between tables 24.6 and 24.7. In **Table 24.6** 'nutritional knowledge' is the **independent variable**, and thus is displayed in the rows, whereas the two groups that are to be compared are displayed in the columns. In **Table 24.7** 'nutritional knowledge' has become the **dependent variable**, influenced by the intervention, and is therefore displayed in columns.

When calculating percentages to be put in the cross-table it is important to remember that the totals for each of the groups which are to be compared should be 100%.

III. CONSTRUCTING CROSS-TABULATIONS APPROPRIATE FOR THE RESEARCH OBJECTIVES

When designing your research project you were asked to produce dummy tables for the data you expected to collect (**Module 13**). These dummy tables were made on the basis of the objectives and the type of study.

Since you now have collected your data and have an idea of their quality and how they can be used, you need to look again in a systematic way at the cross-tabulations to be made.

To construct appropriate cross-tabulations we recommend that you follow the steps below:

1. Review each specific objective and the method chosen for collecting the relevant data.

2. Formulate hypothetical sentences that you consider to be the type of conclusions you expect to reach concerning each objective.

For example, in the cross-sectional survey on malnutrition of <5's, where one of the specific objectives is to determine factors associated with early weaning, expected conclusions could be:

- Mothers who are employed wean their children earlier than mothers who are not employed.

- Mothers who do not attend nutrition talks wean their children earlier than mothers who attend nutrition talks.

The reasons for formulating possible conclusions are that they help you:

- remember the purpose of each tabulation and calculation you undertake;

- avoid wasting time on meaningless calculations and tabulations; and

- keep your data organised so you can more easily write a well-organised report.

3. For each expected conclusion, construct a dummy cross-table that will enable you to derive the right conclusions.

4. Perform the appropriate frequency counts (using the data master sheets) and enter the results in the cells of the cross-table.

5. Interpret the data in the table and write clear conclusions. It is not necessary/advised to describe the contents in each cell of the table as this will bore your readers. (Look at the interpretation of the tables presented earlier in this module)

EXERCISE 1:

Select one specific objective from each of the research projects, formulate expected conclusions and construct the appropriate dummy cross-tables.

IV. GENERAL HINTS WHEN CONSTRUCTING TABLES

- MMake sure that all the categories of the variables presented in the tables have been specified and that they are mutually exclusive (i.e. no overlaps and no gaps) and exhaustive.
- When making cross-tabulations, check that the column and row counts correspond to the frequency counts for each variable.
- Also check that the grand total in the table corresponds to the number of subjects in the sample. If not, an explanation is required. This could be presented as a footnote. (Missing data, for example.)
- Think of a clear title for each table. Also be sure that the headings of rows and columns leave no room for misinterpretation.
- Number your tables and keep them together with the objectives to which they are related. This will assist in organising your report and ensure that work is not duplicated.

V. INTERPRETING RESULTS

All tables presented in this module have been interpreted. You will have noticed that **describing your informants** in terms of profession, location, age or sex is straightforward. You may, however, already notice differences between groups of informants in terms of these background variables, which are interesting to *explore further.* For example, if malnourished children are predominantly from one area, or of the female sex, compared to well-nourished children, you might try to further explore what the underlying reasons are for this difference.

In case of **associations between variables**, it is challenging to identify *in what direction the associations go and what this would mean for the problem under study.* If there is an association between a mother's working status and her weaning practices, while early weaning (= early bottle-feeding) is associated with malnutrition, in particular among lowly educated mothers, one might have a further look at the full-time employed mothers. What is their educational level? It would seem that the non/or lowly educated mothers who are full-time employed might be a specific risk group for malnutrition. This should be checked. Such focussed playing with data is what makes a researcher a 'detective' (see **Module 10B** part V, Interview skills).

Interpretation of data needs to be **correctly phrased.** Particular attention should be paid to the phrasing of conclusions on tables in which **selected groups are compared** with respect to a number of **independent variables**. In this case you have to express the occurrence of the independent variables within the selected groups, and not the other way around.

> **For example**, in **table 24.6** you will state that the mothers of malnourished children appear to have far less knowledge about weaning foods than mothers of well-nourished children. You will **not** state that among the mothers with a low level of knowledge, a higher percentage (82%) was malnourished than well-nourished. This is because you have selected some study populations from your cross-sectional survey you want to compare on the basis of *nutritional status.* Thereby you may have distorted the distribution of mothers with a high or low level of knowledge about weaning foods (those mothers have no equal chance of appearing in your sub-sample).

To help you formulate your results correctly and consistently, it has been stressed in the modules that the groups you compare should each total 100% with respect to the variables you compare them on.

The need for further statistical testing

Your will have noticed that all interpretations of tables have been stated in provisional terms: 'it appears that ...'. Even though there *appears* to be a relationship between variables, or a strong difference between groups that are compared on absence or presence of a certain factor, we can only be sure if this relationship is *confirmed by a statistical test*. **Modules 28-31** will show the statistical procedures required.

Confounding variables

In **Module 8** and **9** we discussed that sometimes there may appear to be a relationship between two variables, but that in reality this relationship is disturbed by another confounding variable, which is related to both.

This is particularly annoying if we try to solve a problem and want to identify which factors are contributing to it. Due to confounding variables we may draw wrong conclusions and therefore take wrong decisions for action.

> **For instance,** in the example given in **Module 8, education** was mentioned as a possible confounding variable, blurring the relationship between bottle feeding practices of mothers and the prevalence of diarrhoea in their children. In **Module 26** it will be shown that by looking only at the relationship between onset of bottle feeding and diarrhoea in the children there was no association: early bottle feeding even seemed to reduce the occurrence of diarrhoea. However, when the mothers were stratified (split up) in two groups according to their level of education, there appeared to be a relationship. Lowly-educated mothers who started bottle feeding early had significantly more diarrhoea in their children under two years old than lowly educated mothers who weaned their children late. For highly educated mothers it made no difference when they started bottle-feeding: the number of diarrhoea episodes in their children was low throughout, because they apparently followed better hygienic practices. *Education was therefore a confounding variable.* The problem among lowly educated mothers would not have become visible if we had not stratified the mothers according to educational level.

When interpreting cross-tables, we always have to be aware of possible confounding variables. In **Module 26** we will discuss the procedures for controlling for confounding variables. Also the analysis of **matched data**, collected to prevent confounding by specific variables such as education or age, will be discussed in Module 26.

GROUP WORK

- Review each specific objective and its relevant research design: formulate hypothetical sentences that describe the type of conclusions you expect for each objective.
- Construct dummy cross-tabulations, keeping in mind whether you want to:
  - describe research subjects in your sample or describe the problem;
  - compare groups in order to find differences; or
  - find associations between variables.

  Refer back to the dummy cross-tabulations which you already made in the first workshop (**Module 13**).
- If you construct analytic cross-tabulations, try to identify possible confounding variables.
- Finally, fill in the dummy cross-tabulations with data, calculate percentages and interpret what they could mean, in relation to your objectives and study questions.

Trainer's Notes

Module 24: CROSS-TABULATION OF DATA

Timing and teaching methods

1 hour	Introduction and discussion
3 hours +	Group work
1 hour	Group presentations and plenary discussion (optional)

Introduction and discussion

- It is recommended that you use an overhead projector (or flipcharts) when presenting and explaining the construction of different tables in order to focus the attention of the participants. Do not merely refer to the modules.

- Some of the cross-tabulations presented in the module are filled in with imaginary data. This was done in order to make examples more concrete and to show how tables should be interpreted. Special attention should be given to the design of the tables (what comes in rows, what comes in columns). Therefore it is recommended that you use 2 transparencies, which can be placed on top of each other, to present the table: one with the dummy table, the other with the data.

- Pay extra attention to how each table should be read: some of them have to be read horizontally, others vertically, depending on whether the groups to be compared are put in rows or in columns.

- The classroom exercise at the end of section III will help to give participants useful practice in designing appropriate cross-tables. Try to obtain at least one table of each of the three types (i.e., descriptive and the two types of analytic tables) from the groups depending on their studies.

- Make sure that there is agreement concerning how to read the table (horizontally or vertically). If groups are having difficulty constructing appropriate tables for their own projects, this indicates that there is a need for the groups to present a few of their cross-tabulations in plenary after the group work session so they can get feed-back.

- The hints on constructing tables (section IV of the module) should be illustrated with an example of a table taken from the module or from one of the groups.

Group work

- Go through the first three steps of the group work assignment with the group as a whole. The data in the tables can be filled in by sub-groups.

- Stress that once the tables are filled in with data they can be interpreted immediately. The tentative conclusions should be recorded right under each table to make the writing of the report during the second week easier. Later these conclusions should be discussed with the other group members.

- At this stage the numbering of tables can be done according to the specific objective to which they relate.

Group presentations and plenary discussion (optional)

- Ask all groups to present at least one table (filled in with data), as well as the conclusions derived for each of their objectives that require cross-tabulations. Other groups and facilitators should be invited to comment.

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 25

MEASURES OF ASSOCIATION BASED ON RISK

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 25: MEASURES OF ASSOCIATION BASED ON RISK

OBJECTIVES

At the end of this session you should be able to:

1. **Define** incidence, risk, relative risk and odds ratio.
2. **Calculate** relative risk for appropriate study designs (cross-sectional comparative studies, cohort studies, case-control studies and experimental studies).

I. **Introduction**

II. **Incidence, risk and relative risk**

III. **Calculating relative risk in different study designs**

I. INTRODUCTION

In HSR studies, the objective of many comparative studies (cross-sectional comparative studies, case-control studies, cohort studies, experimental and quasi-experimental studies) is to compare those who develop the problem/condition under study among those with a risk factor (exposed group) and those without this risk factor (non-exposed group). How this is actually done depends on the study design used.

You will remember (**Module 9**) that in **case-control studies** and cross-sectional comparative studies, one group of study subjects is selected that has the problem under study (cases) and a control group that does not have the problem. The two groups are then compared with regards to the presence or absence of risk factors. In a **cohort study**, only study subjects without the problem to be studied are selected. They are then divided into those with the assumed risk factor and those without the risk factor. The two groups are followed up over a longer period and the occurrence of the problem is then measured between those with the assumed risk factor and those without the risk factor. In an **experimental study**, subjects with a certain problem are selected. One group is then exposed to an intervention by the researcher (epidemiologists call this a 'beneficial risk factor'), while the other group remains untouched by the intervention. After an appropriate follow-up period, the occurrence of the problem is measured and a comparison made between the two groups, in the hope that the intervention has at least partly solved the problem.

The comparison of a group with a risk factor and a group without, or of cases and controls, allows the researcher to determine **whether** the potential risk factor influences the problem or not and **to what extent** the risk factor actually contributes to the problem.

Before one can measure how much a risk factor contributes to a problem, it is important to understand the concept of incidence, risk and relative risk.

II. INCIDENCE, PREVALENCE, RISK AND RELATIVE RISK

Incidence and incidence rate

INCIDENCE is the total number of *new* events or cases of a defined condition (for example a disease) which occur during a specified period of time in a defined population who can develop the condition of interest.

Note: The population who can develop the condition of interest is known as the population 'at risk'.

Example 1:

The total number of new tuberculosis cases in District A in the year 2000 was 273. The **incidence** of tuberculosis in District A in 2000 was therefore 273.

INCIDENCE RATE (cumulative incidence) is the total number of new events or cases of a defined condition that occur during a specified period of time divided by the 'population at risk'.

An incidence rate is usually expressed per 1,000 or per 10,000 or per 100,000 (or other factor of 10) inhabitants to make it easier to compare the rates in different communities.

Example 1 (continued):

District A has a population of 200,000. The **incidence rate** of tuberculosis in the year 2000 in district A was therefore 273/200,000/year or 137/100,000/year. (Divide the numerator and denominator by 2 so the incidence rate is expressed per 100,000 per year, and round off to the nearest whole number, i.e., 136.5 becomes 137.)

The incidence rate estimates the chance (probability or risk) that an individual will develop a disease during a specified period of time.

Note:

PREVALENCE is the total number of *new and old* cases, regardless of when they occurred. The PREVALENCE RATE (old and new cases divided by total population) is very useful in management as it gives an indication of resources needed for dealing with the problem

Example 1 (continued):

If on 31 December 2000, 360 old and new patients would be registered, the prevalence rate would be 360/200,000 or 180/100,000 per year.

Like incidence rates, prevalence rates can be expressed per 1,000, 10,000 or 100,000 inhabitants.

Risk and relative risk

RISK is the same as incidence rate.

Example 1 (continued):

The **risk** of getting tuberculosis in district A in 2000 was 137/100,000/year.

The risk may not be the same for various subgroups in the population. Whereas the risk of getting tuberculosis in farmers might be 100/100,000/year, it may be 200/100,000/year in mine workers. In this example, mine workers were twice as likely to get tuberculosis as farmer.

It may therefore be concluded that being a mine worker is a **risk factor** for contracting tuberculosis and carries a **relative risk** of 2.

A RISK FACTOR is any factor whose presence is associated with an increased risk of a disease or condition.

When determining relative risk we have to consider two subgroups in the study population: the subgroup in which the risk factor is present (exposed) and the one in which the risk factor is absent (unexposed).

RELATIVE RISK is the risk of getting the disease in the group with the risk factor divided by the risk of getting the disease in the group without the risk factor.

$$\text{Relative risk (RR)} = \frac{\text{incidence rate (risk) among those with risk factor (exposed)}}{\text{incidence rate (risk) among those with no risk factor (unexposed)}}$$

Note:

If RR = 1, then the risk of disease is equal in those with the risk factor and those without, resulting in a relative risk of 1. Therefore, there is no association between the risk factor and the problem/condition under study.

If RR > 1, then the risk of disease is greater in those with the risk factor than those without. In this case the factor is associated with the problem/condition under study.

If RR < 1, then the risk of disease is lower in those with the 'risk factor' than those without. The risk factor in this case actually protects against or reduces of the problem. (e.g., a beneficial 'risk factor' such as a Health Education Programme or other intervention, which helps to solve or reduce a certain problem).

It is important to note that the identification of a risk factor does **not** imply that there is a **causal relationship** between the factor and the condition. However, the higher a relative risk is, the more likely it becomes that the risk factor is causal and not due to chance or confounding.

III. CALCULATING RELATIVE RISK IN DIFFERENT STUDY DESIGNS

1. Calculating relative risk in cohort and intervention studies

In cohort and intervention studies, it is possible to calculate the incidence rate (risk) directly. This is because the outcomes or diseases/problems under study occur during the study.

In these studies, the incidence rates in the exposed and non-exposed groups are calculated, which are then used to calculate the **relative risk** using the formula presented in section II of this Module.

In a cohort study, the data needs to be put in the table format given in **Table 25.1**.

Table 25.1. General cohort study table format for risk calculation

Presence of the risk factor	Presence of the problem		TOTAL
	Yes	No	
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	a + b + c + d

Where: a = Subjects with the risk factor who also have the problem under study
b = Subjects with the risk factor who do not have the problem
c = Subjects without the risk factor who have the problem
d = Subjects without the risk factor who do not have the problem

From this: a+b = all subjects who have the risk factor
c+d = all subjects without the risk factor

The risk of developing the problem among those with the risk factor = $a/(a+b)$

Risk of developing the problem in those without the risk factor = $c/(c+d)$

The **relative risk** will then be equal to $\frac{a/(a+b)}{c/(c+d)} = \frac{a/(c+d)}{c/(a+b)}$

Example 2:

A study was carried out in country X to find out whether the risk of diarrhoea in under-five children was different between two nearby wards (each made up of a number of villages). It was suspected that Ward B had a higher risk because the community used unprotected wells while Ward A used borehole water. Ward A had a population of 10,000, while Ward B included 15,000 people. The respective under-five populations were 1,000 and 1,500. The records kept by village health workers (VHWs) responsible for the wards in the previous four weeks were checked. Ward B had had 78 cases of diarrhea while Ward A had had 50 cases.

Is the risk of diarrhea different between the two wards?

Because Ward B had the potential risk factor (unprotected water sources), the population of this area is expected to be at higher risk of developing diarrhoea. This data can now be put in a cross-tabulation using the format of table 25.1, and the **risk** and **relative risk** can then be calculated using the formulae already discussed.

Table 25.2. Relationship between exposure to unsafe water and diarrhea in under-5 children

Exposure	Presence of diarrhea		TOTAL
	Diarrhea	No diarrhea	
Unsafe water (Ward B)	78 (5%)	1,422 (95%)	1,500 (100%)
Safe water (Ward A)	50 (5%)	950 (95%)	1,000 (100%)
Total	128 (5%)	2,372 (95%)	2,500 (100%)

Combine **Table 25.1** and **25.2**: a = 78; b = 1,422; c = 50; d = 950
a+b = 1,500; c+d = 1,000

Risk of developing diarrhea in Ward B (unsafe water) = $a/(a+b)$ = 78/1,500

Risk of developing diarrhea in Ward A (without the risk factor) = $c/(c+d)$ = 50/1,000

The **relative risk** will then be equal to $\frac{78/1,500}{50/1,000} = \frac{78 \times 1,000}{50 \times 1,500} = 1.04$

The 95% confidence intervals (CI) for this RR are calculated for you, and are 0.74 to 1.47.

Interpretation:

Under-five children in Ward B have < 1.04 times more risk of developing diarrhea than those in Ward A, but this risk is not significantly different as the 95% CI includes 1. (This significance will be further evaluated using statistical significance tests in **Modules 28-30**.)

Note:

The calculation of the relative risk for experimental, quasi-experimental and cross-sectional comparative study designs based on *incidence* data is exactly the same as for cohort studies. (See **Annex 25.1** for examples of the calculation of relative risks for these designs.)

2. Estimating relative risk in a case-control study

In a case-control study, it is usually not possible to calculate relative risk directly as in the incident studies discussed earlier.* In most situations, the odds ratio (OR) is used, which *estimates* the relative risk.

To compute the OR, the data has to be presented in the table format given in table 25.3.

Table 25.3: Case-control study table format

Risk factor	Cases	Controls	TOTAL
Present (+)	a	b	a+b
Absent (-)	c	d	c+d
Total	a+c	b+d	a+b+c+d

Where:
- a = Subjects with the problem (cases) who have the risk factor
- b = Subjects without the problem (controls) who have the risk factor
- c = Subjects with the problem (cases) who do not have the risk factor
- d = Subjects without the problem (controls) who do not have the risk factor

$$OR = \frac{ad}{bc}$$

The OR is sometimes called the *cross-products ratio*. It is the product of the left upper cell (a) and the right lower cell (d) or a times d, divided by the product of the right upper cell (b) and the left lower cell (c), or b times d.

After calculating the odds ratio (OR), we usually call it the relative risk (RR) as that is what it is estimating.

* In 'incident studies' the incidence of a certain problem or outcome and the risk or contributing factor can both be observed; therefore the relative risk can be calculated directly.

Example 3:

An HSR case-control study was conducted in Namibia to identify factors contributing to early **neonatal mortality** (first seven days of life) in the maternity hospital in the capital Windhoek (Muharukua et al. 1998). For each case, 5 controls were selected. The final sample size was 290, 49 of whom were neonates who died between birth and day seven of life (cases) and 241 neonates who survived the first seven days of life (controls). Among the potential risk factors evaluated was low birth weight (less than 2,500 g).

Of the 281 neonates about whom information was available on low birth weight (LBW), 44 were cases and 237 controls. 28 of the 44 cases were LBW, while 29 of the 237 controls had a low weight.

This data, put in the cross-tabulation format given in table 25.3, is shown in **Table 25.4.**

Table 25.4: Relationship between prematurely and neonatal death in a case-control study in Namibia

Low birth weight	Neonatal deaths	Controls	TOTAL
Yes	28 (64%)	29 (12%)	57
No	16 (36%)	208 (88%)	224
Total	44 (100%)	237 (100%)	281

Where a = 28; b = 29; c = 16; d = 208

The formula for odds ratio (OR) is: $\frac{ad}{bc}$

Therefore $OR = RR = \frac{28 \times 208}{29 \times 11} = 12.55$

The 95% CI, (calculated using Epitable in Epi Info 6.04c as shown in the Computer Companion, Vol II Part 3 of the HSR Training Series, WHO/AFRO, Harare, 1996) is 5.69 - 28.00 (see **Module 27** for formulae for calculating 95% CI).

Interpretation:

The risk of a neonate born with a low birth weight in the maternity hospital in Windhoek dying in the first seven days of life is 12.55 times that of normal weight neonates. Low birth weight is therefore a very strong risk factor for neonatal death.

Note:

That the percentages of cross-tabulations already gave you a clue that in **Table 25.2** the OR/RR would be low, whereas in **Table 25.4** you could assume it would be high. These two examples are provided for teaching purposes. In reality you would not even bother to calculate the odds ratio for table 25.2.

3. Calculating relative risk in a cross-sectional comparative study (prevalence survey)

As in a case-control study, incidence will usually not be directly calculated in a cross-sectional comparative *prevalence* study. The measure of association in these studies is called a prevalence odds ratio (POR). The POR is calculated in exactly the same way as the odds ratio in a case-control study.

Example 4:

In Botswana, a cross-sectional comparative study was conducted to determine the magnitude of the problem of **teenage pregnancy** and to identify contributing factors. The researchers sampled 400 teenagers at random and found that 23% of the teenagers had been pregnant. Among other things, they wanted to evaluate whether teenagers who had received organisational support (i.e., peer education) would be less likely to become pregnant than those who had not received support.

From the collected data, only 14 of the 90 teenagers who had experienced a pregnancy had received organisational support while 86 of the 310 never pregnant teenagers had received support. This data is shown in **Table 25.5.**

Table 25.5: Relationship between organisational support and teenage pregnancy in Botswana

Organisational support	Experienced pregnancy		TOTAL
	Yes	No	
No	76 (25%)	224 (75%)	300 (100%)
Yes	14 (14%)	86 (86%)	100 (100%)
Total	90 (23%)	310 (77%)	400 (100%)

Where a = 76; b = 224; c = 14; d = 86

(Prevalence) odds ratio (OR) $= \frac{a \times d}{b \times c}$

OR = RR $= \frac{76 \times 86}{224 \times 14} = 2.08$

The 95% CI is 1.07 to 4.11 (calculated from Epitable in Epi Info version 6.04 c, Computer Companion, Vol II Part 3 of HSR Training Series).

Interpretation:

Teenagers who received no organisational support were 2 times more likely to become pregnant than those who received support. Please note that this finding is both practically and statistically significant.

Warning

When you find an association between a potential risk factor and the problem, this may not be an actual relationship. There may be a number of reasons for this finding, one of which is that there could be other risk factors (confounders) that provide the actual explanation. These other risk factors therefore need to be taken into account. **Module 26** deals with the issue of confounding, including how to evaluate and control for it.

GROUP WORK

If you performed a comparative study (cross-sectional comparative study, cohort study, case-control study or (quasi)-experimental study):

- Analyse your two-by-two cross-tables and select tables where the percentages indicate that there could be an association between two variables.
- Calculate the Odds Ratio/RR for the tables.
- Interpret the results and write the interpretation under each table

Annex 25.1: Relative risk calculations for quasi-experimental and experimental study designs, and for cross-sectional comparative studies using incidence data.

The cohort study on the association between exposure to unsafe water and diarrhoea in under-5 children shown in example 2, was actually done as the first part of a quasi-experimental intervention study. The study showed that the occurrence of diarrhoea was the same for Ward A and Ward B.

Example 2 (continued):

The researchers then developed a health education intervention package and proceeded to carry out an intervention study to evaluate the effectiveness of the package. The health education intervention was implemented in Ward A over a period of 2 months, while the usual health services were available in Ward B. The researchers were confident that any difference in risk of diarrhea after the intervention could be attributed to it, as the incidence in wards A and B was the same before the intervention. After the 2 months, VHWs again recorded the incidence of diarrhea over a period of 4 weeks in the two wards. Ward A recorded 16 cases while Ward B recorded 55 cases.

The study design used here is a quasi-experimental study, because there is comparison but no randomisation (see **Module 9**, intervention studies).

The question driving the study was: Does the intervention work; does it result in a significant decrease in diarrhoea?

Table 25.6: Relationship between exposure to unsafe water and diarrhoea in under-5 children after the intervention

Exposure	Diarrhea	No diarrhea	TOTAL
No intervention (Ward B)	55	1445	1,500
Intervention (Ward A)	16	984	1,000
Total	71	2,429	2,500

From table 25.6, a = 55; b = 1445; c = 16; d = 984

Using the formula presented after table 25.3: $a+b = 1{,}500$, $c+d = 1{,}000$

Risk of developing diarrhoea in Ward B (no intervention) = $a/(a+b) = 55/1{,}500$

Risk of developing diarrhoea in ward A (intervention) = $c/(c+d) = 16/1{,}000$

The **relative risk** will then be equal to $\frac{55/1{,}500}{16/1{,}000} = \frac{55 \times 1{,}000}{16 \times 1{,}500} = 2.29$

The 95% confidence interval (CI) for this RR is calculated for you, and is 1.64 to 3.98 (using Epitable of Epi Info version 6.04c, see Computer Companion, Vol. II Part 3 of the HSR Training Series).

Interpretation:

The health education package does work. The risk of diarrhea in under-fives after the intervention is now much higher (2.29 times) in Ward B than in ward A. The finding is both practically important (the risk of diarrhoea is twice as high in the non-intervention group) and statistically significant (95% confidence levels do not include 1). Formal statistical testing will be covered in **Modules 29 and 30**. Remember that the risks of diarrhea were not significantly different before the intervention (**Table 25.2**).

Calculating relative risk in a cross-sectional comparative study (incidence study)

The incidence rate can also be calculated directly in a cross-sectional comparative study where incident cases (problems) are identified. This can then be used to calculate the relative risk as already shown (**Table 25.2**).

The data however has to be put in a table format as given in **table 25.7**.

Table 25.7: Cross-sectional study table format (incident outcomes)

Risk factor	Incident Yes	Incident No	TOTAL
Present (+)	a	b	a+b
Absent (-)	c	d	c+d
Total	a+c	b+d	a+b+c+d

Where: a = Subjects with the risk factor who also have the problem or incident under study
b = Subjects with the risk factor who do not have the problem
c = Subjects without the risk factor who have the problem
d = Subjects without the risk factor who do not have the problem

The formula given for calculating relative risk can then be applied.

Incidence in the exposed = a/(a+b)

Incidence in the non-exposed = c/(c+d)

Relative risk can also be written as:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a \times (c+d)}{c \times (a+b)}$$

Example 5

In a study carried out on factors that contribute to **delayed perinatal care** in Machinga District, Malawi in 1997 (Tamaona et al, 1997), 88% of the 97 ANC mothers interviewed delayed seeking perinatal care (until after 3 months of pregnancy). To find out whether distance from health facilities was a risk factor for this delay in perinatal care, the mothers were separated into those living 10 km or less from health facility (near) and those living more than 10 km away (far). 60 of the 62 mothers living far from the health facility delayed using perinatal care compared to 28 of 35 mothers living near.

The data is shown in **Table 25.8** using the format given in **Table 25.7.** The incidence (problem) is delay in seeking perinatal care while the risk factor is living far (>10 km).

Table 25.8: Relationship between distance and delay in seeking perinatal care in Machinga District, Malawi

Distance from Health facility	Seeking perinatal care		TOTAL
	Delay	No delay	
> 10 km	60	2	62
0 -10 km	28	7	35
Total	88	9	97

Of mothers living far, 96.77% delayed. This is the same as

$$60/62 = 96.77/100 = 0.9677$$

On the other hand, 80.0% of those living near delayed, i.e.

$$28/35 = 80.00/100 = 0.80$$

$$RR = \frac{0.9677}{0.80} = 1.21$$

The 95% CI for this RR is 1.02 - 1.44 (worked out using Epitable in Epi Info Version 6.04c, refer to Computer Companion, Volume II Part 3 of the HSR Training Series.)

Interpretation:

Women using ANC who live beyond 10 km from a health facility in Machinga District, Malawi, are therefore 1.21 times more likely to delay presenting for perinatal care than those living 10 km or less from the nearest health facility. The difference is statistically significant, but it is not a very important one (only 1.2 times more likely to delay). Improving access here may not change delay in perinatal care that much. There must be other factors, overriding the distance, which make mothers stay away.

Trainer's Notes

Module 25: MEASURES OF RISK AND ASSOCIATION

Timing and teaching methods

1½ hour	Introduction and discussion
2 hours	Group work

Introduction and discussion

It is important that the participants understand the concept of risk. Try to ensure that they also have a good idea of how to apply this concept in all study designs. Let them calculate some odd ratio's in plenary on their own data.

Impress on them the importance of quasi-experimental study designs and how these can be used to evaluate programmes in a scientific manner.

The team of facilitators should provide computer support to improve on depth of analysis as manual data analysis may result in inadequate analysis of the contribution of the risk factors to the problem under study.

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 26

DEALING WITH CONFOUNDING VARIABLES

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 26: DEALING WITH CONFOUNDING VARIABLES

OBJECTIVES

At the end of this session you should be able to:

1. **Explain** different ways of dealing with confounding at the design and analysis stage of a study.
2. **Evaluate** whether an association between two variables may be influenced by another confounding variable/risk factor.
3. **Calculate** association in a way that takes into consideration the effect of potential confounding by another variable/risk factor.

I. Introduction

II. The nature of confounding

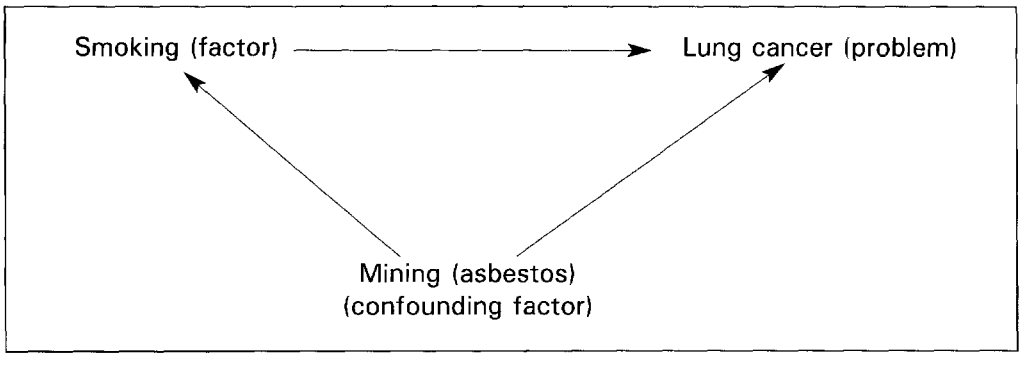
III. Methods to control confounding in the analysis

I. INTRODUCTION

In **Module 25**, we calculated risk and then measured the association between the risk factor(s) and the problem under study (outcome). However, it is possible that the observed association is partially or totally due to another risk factor, not the one being considered. If this 'distortion' being introduced by this second risk factor is not taken into consideration, we will not be able to tell if the first association actually exists or is as strong as we calculated.

For example (1), it is known that smoking is a risk factor for lung cancer. However, in region Z, many men worked in the local asbestos mines. They were therefore exposed to asbestos, which is a known risk for lung cancer. It is also known that, because of the stress miners are exposed to, they tend to smoke more, especially when working underground. Whereas smoking is related to lung cancer, mining is related to smoking as well as to lung cancer. Therefore, there is a triangular relationship between smoking, mining and lung cancer, as shown in **figure 26.1**.

Fig. 26.1: Inter-relationship between smoking (factor), mining (confounding factor) and lung cancer (problem) in a cohort study



This module will discuss how the influence of confounding variables can be removed, so that it will be possible to calculate how much the original risk factor on which a researcher wants to concentrate (smoking in example 1) contributes to the problem under investigation (lung cancer).

II. THE NATURE OF CONFOUNDING

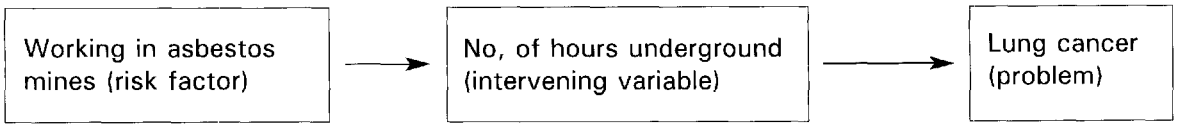
What is a confounding factor?

A CONFOUNDING FACTOR is an independent variable that distorts the association between another independent variable and the problem under study, as it is related to both.

For a variable to be confounding, it must be associated with the first risk factor (smoking in example 1) and be an independent risk factor for the problem (lung cancer). On the other hand: a variable, which is merely a link between the risk factor we are looking for and the problem, is not a potential confounder.

For example, if we would look at the number of hours miners are underground in relation to lung cancer, (see **Figure 26.2**) this would just be a further specification of the risk factor working in the asbestos mines. Therefore it would not blur the association between working in the asbestos mines and lung cancer. (See also the example of **Module 23**, Section III.)

Fig. 26.2: Inter-relationship between a factor, an intervening variable and the problem or outcome



For a factor to be a potential confounding variable there has to be a **triangular** relationship between the first risk factor, the potential confounding factor and the problem under investigation, as shown in **Figure 26.1**.

What is the effect of confounding?

Confounding can result in the association between a risk factor and the outcome appearing smaller (under-estimated) or appearing bigger than it is (over-estimated). It can even change the direction of the observed effect, resulting in a harmful factor appearing to be protective or vice versa.

Let us reconsider the effect of working in the asbestos mines on the relationship between smoking and lung cancer.

Example 1 (continued)

The risk of a smoker developing lung cancer was found to be 10 times that of a non-smoker in region Z. When the researchers removed the confounding effect of working in the asbestos mines, they found that smokers were only 7 times more likely to develop lung cancer. The true relationship between smoking and lung cancer was therefore 30% lower than when the confounder would not have been dealt with.

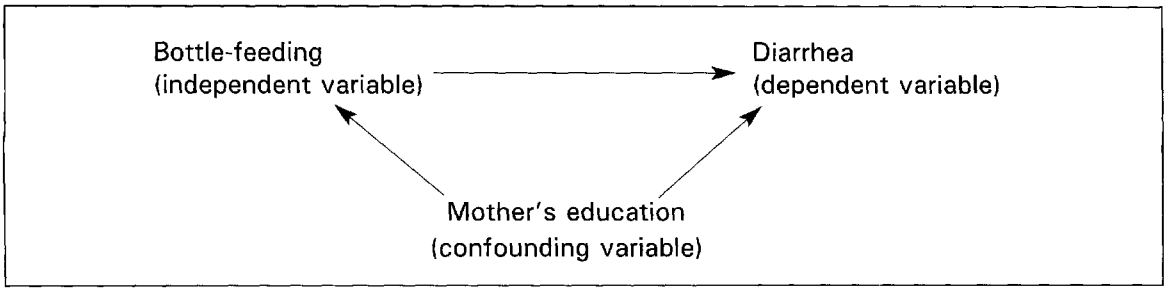
In **Module 8**, where we introduced the concept of confounding variables, we gave an example of how the effect of a mother's practice of bottle-feeding on her child suffering from diarrhoea could be confounded by the mother's educational level. We will now remove the effect of the mother's education, in order to be able to discover the true relationship between bottle-feeding and diarrhoea.

Example 2:

In a cross-sectional comparative study using prevalent cases of infants ever having had diarrhoea (discussed in **Module 24** Section V), it was assumed that bottle-fed infants are more likely to develop diarrhoea than breast-fed infants. It is known that uneducated mothers are less likely to bottle-feed their infants but their infants are still more likely to have diarrhoea than those of their more educated counterparts. Mothers' education is therefore a possible confounder of the relationship between bottle-feeding and diarrhoea (see **Figure 26.3**).

In this study, it was found that bottle-fed infants were less likely (0.8 times or OR = 0.8) to develop diarrhoea than breast-fed ones. This suggested that bottle-feeding *protected* infants from diarrhoea! However, when mother's education was taken into account, it was found that bottle-fed infants were actually 1.2 times more likely to develop diarrhoea than breast-fed infants. This association had been distorted by the effect of maternal education. The true relationship between bottle-feeding and diarrhoea was actually 50% larger than the distorted (confounded) one.

Fig. 26.3: Inter-relationship between bottle-feeding (risk), mother's education (confounding factor) and diarrhoea (problem) in a case control study



How to identify confounding factors

Because confounding factors can distort true relationships between other risk factors and the problem under study, it is critical that they are considered during the *design of data analysis*. To be able to control for confounding variables at the data analysis stage, data on the confounder has to be collected during data collection. Otherwise the distortion due to confounding factor(s) cannot be removed.

The problem analysis diagram (**Module 4**) provides insight into some possible confounding variables, as it visualises the triangular relationships just discussed. Known confounding variables from literature should be taken into consideration, while others can be identified when brain storming about possible risk factors and their inter-relationships. **Background variables** such as age, education, economic status are notorious confounders, as they are related to many aspects of life (see **Module 8**). They are even often inter-related, e.g., education and age may both contribute to economic status.

During data analysis, other potential confounding factors can be identified when unusual or unexpected findings are observed. When relationships are much stronger, weaker or even opposite (e.g., protective instead of being harmful) of what would be expected from what is known from the literature and your own experience, this should alert you to look for possible confounders.

EXERCISE

Use your problem analysis diagram and the findings you have produced up to now to identify possible confounding factors.

III. METHODS TO CONTROL FOR CONFOUNDING IN THE DATA ANALYSIS

Control of confounding in data analysis is achieved by stratified analysis or by multivariate analysis. In a stratified analysis, the objective is to compare the group of cases and of controls within homogeneous categories of the confounding variable. Multivariate analysis, on the other hand, involves the use of a mathematical model and allows the researcher to control for all confounders at the same time while looking at the contribution of each risk factor to the problem. Multivariate analysis will not be discussed any further, but those interested in using it are referred to epidemiological or statistical handbooks. (See references in **Modules 9** and **28**.). When researchers when designing the study expect that some variables could be confounding the relationship between the problem under study and a contributing factor, they may as well have matched the cases and controls for these 'possible confounders'. Matched data require specific analysis (see part 2 of this section).

1. Stratified analysis

The steps in controlling for confounding through stratified analysis are:

1. Calculate the relative risk (RR) or odds ratio (OR, which is an estimate of the RR, see **Module 25**) without stratifying (crude RR or crude OR)
2. Stratify by the confounding variable
3. Calculate the adjusted RR (or OR)
4. Compare the crude RR or OR with the adjusted RR or OR

If the adjusted estimate (aRR or aOR) is *equal* to the unadjusted one (RR or OR), then there is no confounding. If they are *different*, then there is confounding. But one may ask, how big should the difference be? Epidemiologists have generally agreed that:

> If the CRUDE RELATIVE RISK differs from the ADJUSTED RELATIVE RISK by 10% or more, there is important confounding. The adjusted RR should then be calculated by stratifying the confounding variable.

The 95% CI (and formal significance testing) can now be carried out to measure the significance of the association between the risk factor and the problem for the different strata.

(1) Stratified analysis to control for confounding in an incident study*

The procedure for controlling for confounding is the same in all incident studies, including case-control and cohort studies, cross-sectional comparative studies and experimental studies.

Example 3:

In the intervention study presented **Module 25** (example 2), age was suspected to be a confounding factor. Children in Ward A were found to be generally older than those in Ward B (less than 2 years vs. 2 years and above) and older children are more likely to develop diarrhoea. We therefore need to assess if age confounds the relationship between the health education intervention and the development of diarrhoea, and split up **Table 25.2** for children <2 years and children $\geq$ 2 years.

Of the 1500 children in Ward B, 840 were < 2 years and 660 were 2 years or older. Among the 1000 children in Ward A, 210 were < 2 years and 790 were / 2.

Of the 840 children younger than two years in Ward B, 40 had diarrhea compared to 10 of the 210 children younger than two in Ward A. Among the children 2 years old and older, 15 of 660 had diarrhoea in Ward B compared to 6 of 790 in Ward A (see **Tables 26.1** and **26.2**).

These tables now show the data *stratified by age* to assess for confounding. It had already been found in **Module 25** Annex 25.1, example 2 that the crude RR was 2.29, i.e., the risk of diarrhea in the intervention ward was just over twice that in the non-intervention ward, suggesting that the health education package actually worked. The adjusted RR is 1.51. (See **Annex 26.1** for formulae. Epitable in Epi Info 6.04 can also be used to do the calculations, see Computer Companion, HSR Vol. II Part 3.)

* In 'incident studies' the incidence of a certain problem or outcome and the risk or contributing factor can both be observed; therefore the relative risk can be calculated directly.

Table 26.1: Risk of diarrhoea in children under 2 years by intervention area (post intervention)

Exposure	Diarrhea	No diarrhea	TOTAL
No intervention (Ward B)	40	800	840
Intervention (Ward A)	10	200	210
Total	50	1,000	1,050

Table 26.2: Risk of diarrhoea in children 2 years or older by intervention area (post intervention)

Exposure	Diarrhea	No diarrhea	TOTAL
No intervention (Ward B)	15	645	660
Intervention (Ward A)	6	784	790
Total	21	1,429	1,450

Is the crude RR (RR) different from the adjusted RR (aRR)?

2.29 is different from 1.51.

Is this difference greater than 10%?

RR - aRR = 2.29 - 1.51 = 0.78

The difference between RR and aRR as a percentage of RR is therefore:

$$\frac{0.78 \times 100}{2.29} = 34\%$$

Since the aRR differs from the RR by over 10% (actually it is 34%), there is considerable confounding. The aRR should therefore be used, as the RR is distorted by the effect of age.

The 95% CI from Epitable (Epi Info 6.04c) is 0.89 - 2.55.

Interpretation:

Age confounded (distorted) the relationship between the intervention and the occurrence of diarrhea, giving an over-estimate of the RR of 2.29 when the true relationship was only 1.51. The *intervention carried out in Ward A does not seem to have worked* as, even though the risk of diarrhea among the intervention Ward under-fives is 1.51 times that of the non-intervention Ward (B), this could also occur by chance (95% CI includes 1).

(2) Stratified analysis to control for confounding in a case-control study

The principle of controlling for confounding using stratification is the same for case-control and cross-sectional comparative studies using prevalent outcomes, as they both use the Odds Ratio (Prevalence Odds Ratio in the latter) for estimating relative risk.

Example 4:

In the case-control study carried out in Windhoek hospital, Namibia to find factors contributing to early neonatal mortality already referred to in **Module 25** example 3 (Muharukua et al, 1998), it was suspected that prematurity confounded the relationship between low birth weight (LBW) and neonatal death.

If you think it through, a premature neonate is more likely to die than a full term one, and a premature neonate is also likely to weigh less than 2,500 g (LBW). It is on these grounds that it was suspected that prematurity might have been confounding the relationship between LBW and neonatal death.

It had already been found that the low birth weight neonates were 12.55 times more likely to die than normal weight new-borns. This finding was statistically significant (95% CI is 5.69 - 28.00). It is however possible that this was not the true relationship, as prematurity may have been confounding (distorting) it. We therefore need to stratify the data by the possible confounder, prematurity. Below are the data that can be used to get the two strata.

Of the 281 neonates with data on prematurity and birth weight, 50 were premature while 231 were full term. Among the premature, 27 of the 28 babies that died were LBW compared to 17 of the 22 controls. For the full term neonates, 1 out of the 16 babies that died were LBW compared to 12 of the 215 controls. **Tables 26.3** and **26.4** give the *stratification by prematurity*, using this data.

Table 26.3: Relationship between low birth weight and neonatal death among *premature neonates* in Windhoek hospital

Low birth weight	Neonatal deaths	Controls	TOTAL
Yes	27	17	44
No	1	5	6
Total	28	22	50

Table 26.4: Relationship between low birth weight and neonatal death among *full term neonates* in Windhoek hospital

Low birth weight	Neonatal deaths	Controls	TOTAL
Yes	1	12	13
No	15	203	218
Total	16	215	231

The adjusted OR (aOR) is 3.20 (calculated using the formula in **Annex 26.1** or Epitable of Epi Info 6.04c).

We must now ask, if the OR is different from the aOR

Yes it is (12.55 vs. 3.20!).

By how much?

The adjusted relationship of aOR = 3.20 is 75% smaller than the confounded one (OR = 12.55).

Interpretation:

Prematurity is a very strong confounder of the relationship between low birth weight and neonatal death. If the confounding due to prematurity had not been taken into consideration, the researchers might have concluded that low birth weight increased the risk of early neonatal death 12 times, an apparently very important practical and statistical significant finding. This is, however, not true, as most of this high risk was not due to low birth weight per se. Removing the contribution of prematurity (controlling for confounding by prematurity) shows that low birth weight increased the risk of a neonate dying only 3 fold (aOR=3.20; 95% CI is 0.88 - 11.56), a finding that is actually not statistically significant (95% CI includes 1).

One can therefore correctly conclude that *low birth weight on its own is not a strong risk factor for neonatal death,* but it works mainly through *prematurity. It is therefore prematurity which, if dealt with, will reduce the incidence of neonatal mortality.*

Alternative explanations for non-significant findings

If confounding has been excluded as a possible explanation for results and one finds an association which is not statistically significant (using confidence intervals or formal significance testing, see **Modules 29** and **30**), there are two alternative explanations.

i. The study did not have the power to show the difference even if it exists in the population. This means that the subjects were too few. One can see this from the *confidence intervals*. If the CIs are *wide* (e.g. 0.8 - 50), this suggests a power problem. Another study may need to be conducted to exclude the possibility of confounding. One should try to prevent this problem during proposal development by calculating a sample size that ensures adequate numbers for the study (see **Module 11**).

ii. This is a true finding and there is actually no association or relationship between the suspected factor and the problem. *Narrow confidence intervals* (e.g. 0.85 - 4.55) will indicate that this might be the case.

2. Matched analysis

In some case-control studies, a control is matched to each case to deal with confounding. Matching has to be included in the study design. In the analysis we have to take account of this design. However, if a frequency matched case-control study has been carried out, then a stratified analysis could be done instead, as was done in **tables 26.3** and **26.4.**

Calculating relative risk (OR) in matched case-control studies

In a matched case-control analysis, the important consideration is the NUMBER of PAIRS and not the actual number of individuals in the study.

Example:

A case-control study was carried out to determine causes of a cholera outbreak in one of the slums of Bombay, India. For each cholera case (bacteriologically confirmed) a control was sought who was of the same sex, born in the same age decade, and living in the same neighbourhood. **Table 26.5** presents the results:

Table 26.5: Source of drinking water by cholera patient/control pairs in the 5 days preceding illness

Healthy controls	Cholera cases		Total No. of pairs
	Shallow well	Tap water	
Shallow well	12	3	15
Tap water	30	31	61
Total	42	34	76

It is found that in 30 pairs, the cases drew their water from shallow wells, controls did not, whereas in only 3 pairs, controls drew their water from the shallow wells while cases did not.

Therefore, the relative risk is estimated as:

$$\text{Relative risk} = \frac{30}{3} = 10$$

In other words, those who drew water from the shallow wells were 10 times more likely to get cholera than those who drew water from the taps.

The steps for performing this general type of analysis are described below:

Step 1. Prepare a table:

Distribute the pairs in the 4 cells of the table according to whether

a) both the case and the control have the risk factor (q)
b) the control has the risk factor and the case does not (r)
c) the control does not have the risk factor while the case does (s)
d) both the control and the case do not have the risk factor (t)

Controls	Cases	
	Risk Factor (+)	**Risk factor (-)**
Risk factor (+)	q	r
Risk factor (-)	s	t

Step 2. Examine whether an association exists between the risk factor and the disease or condition. This is done by comparing r and s in the table. Perform a McNemar's X^2 test to determine whether the association is statistically significant, i.e., is not due to sampling variation (see **Module 30**).

Step 3. Estimate the relative risk by using the following formula:

$$\text{Relative risk} = \frac{s}{r} = \frac{30}{3} = 10$$

GROUP WORK

If you performed a comparative study (cross-sectional comparative study, cohort study, case- control study or (quasi)-experimental study):

- Identify and list all the potential confounding variables and the relationships they may be confounding.
- Use stratification to control for the confounding on all the appropriate relationships when needed (if you are not familiar with Epi Info, your facilitator can help you getting the printouts from Epitable).
- Now re-interpret each of the relationships, writing a statement for each of them.

Annex 26.1: Formulae for calculating adjusted relative risks

For all the formulae, T = a+b+c+d = total of all study subjects for the stratum or level.

Cohort study:

$$aRR = \frac{\sum(a \times \{c+d\})/T}{\sum(c \times \{a+b\})/T)}$$

For the stratified data in example (found in **tables 26.1** and **26.2**):

$$aRR = \frac{[(40 \times \{10+200\})/1{,}050] + [(15 \times \{6 + 784\})/1{,}450]}{(10 \times \{40 + 800\})/1{,}050] + [(6 \times \{15 + 645\})/1{,}450]}$$

$$= \frac{(40 \times 210)/1{,}050] + [(15 \times 790)/1{,}450]}{[(10 \times 840)/1{,}050] + [(6 \times 660)/1{,}450]}$$

$$= \frac{8 + 8.172}{8 + 2{,}731} = \frac{16.172}{10.731} = 1.51$$

Note that this is the same answer obtained by using Epitable in Epi Info version 6.04.

Case-control study (**tables 26.3** and **26.4**):

$$aOR = \frac{\sum(ad/T)}{\sum(bc/T)}$$

For example,

$$aOR = \frac{(27 \times 5)/50 + (1 \times 203)/231}{(17 \times 1)/50 + (12 \times 15)/231} = \frac{135/50 + 203/231}{17/50 + 180/231}$$

$$= \frac{2.70 + 0.879}{8 + 0.779} = \frac{3.579}{1.119} = 3.20$$

Note again that this answer totally agrees with that calculated using Epitable in Epi Info version 6.04.

Trainer's Notes

Module 26: DEALING WITH CONFOUNDING IN RESEARCH

Timing and teaching methods

1½ hour	Introduction and discussion
2 hours (+ or -)	Group work

Introduction and discussion

It is more important that participants understand the concept of confounding and know how to identify potential confounding factors than that they fully understand how to use the formulae for determining the effects of confounding.

Facilitators should assist the participants in calculating the adjusted relative risks, preferably by computer.

It is recommended that at least one facilitator is comfortable with the use of Epi Info, especially Epitable. Only version 6.04a or a later version should be used, as the earlier versions have problems in the Epitable calculations. The HSR Computer Companion (Vol. II Part 2 of the HSR Training Series, Harare, WHO/AFRO HSR Unit, 1996) should be available during the course to be used as a reference.

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 27

PREPARATION FOR STATISTICAL ANALYSIS:
Measures of dispersion, normal distribution and sample variation

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 27: PREPARATION FOR STATISTICAL ANALYSIS

OBJECTIVES

At the end of this session you should be able to:

1. **Explain** what is meant by a range, a percentile, a standard deviation, a normal distribution, a standard error and a 95% confidence interval.
2. **Calculate** ranges, standard deviations, standard errors and 95% confidence intervals for your own data, where appropriate.
3. **Interpret** the results of these calculations.

I. Introduction

II. Measures of dispersion

III. The normal distribution

IV. The relation of the sample to the whole study population

V. How to determine the extent to which the findings in the sample are representative for the whole population

I. INTRODUCTION

As we have seen in **Module 22**, the mean, median and mode are measures of the central tendency of a variable but they do not provide any information of how much the measurements vary or are spread. This module will describe some common measures of variation (or variability), which in statistical textbooks are often referred to as measures of dispersion. Furthermore, the concepts of normal distribution, standard error and confidence interval will be introduced. We will need these concepts later in **Modules 28 to 31**, when we will carry out statistical tests.

II. MEASURES OF DISPERSION

1. Range

The RANGE of a set of measurements is the difference between the smallest and the largest measurement.

For example, if the weights of 7 pregnant women were 40, 41, 42, 43, 44, 47 and 72 kg, the range would be 72 - 40 = 32 kg.

Although simple to calculate, the range does not tell us anything about the distribution of the values between the two extreme ones.

If the weights of 7 other pregnant women were 40, 46, 50, 55, 60, 65 and 72 kg the range would also be 72 - 40 = 32 kg, although the values are very different from those of the previous example.

2. Percentiles

A second way of describing the variation or dispersion of a set of measurements is to divide the distribution into percentiles. As a matter of fact the concept of percentiles is just an extension of the concept of the median, which may also be called the 50th percentile.

PERCENTILES are points that divide all the measurements into 100 equal parts.

The 3rd percentile (P3) is the value below which 3% of the measurements lie.

The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie.

To determine percentiles, the observations should be first listed from the lowest to the highest just like when finding the median.

The concept of percentiles is used by nutritionists to develop standard growth charts for specific countries from a representative sample of children whose weight and height are measured according to their age in months.

3. Standard deviation

To determine how much our measurements differ from the mean value there is a measure that we use when applying statistical tests. This measure is called the standard deviation.

> The STANDARD DEVIATION is a measure that describes how much individual measurements differ, on the average, from the mean.

To obtain the standard deviation of a set of measurements you have to complete the following steps:

1. Calculate the mean of all the measurements.
2. Calculate the difference between each individual measurement and the mean.
3. Square all these differences.
4. Take the sum of all squared differences.
5. Divide this sum by the number of measurements minus one.
6. Finally (since the differences from the mean have been squared), take the square root of the value obtained (in order to get back to the same unit of measurement).

Example 1:

11 children of 3 years of age were weighed and the following weights were obtained:

13, 14, 14, 15, 16, 16, 16, 17, 17, 18 and 20 kg

The number of measurements (n) is 11.

To calculate the standard deviation:

(1) We first calculate the mean: the mean value is 16 kg.

(2) Next, we calculate the distance of each measurement from the mean (16-13 = 3; 16-14 = 2; etc.):

3, 2, 2, 1, 0, 0, 0, 1, 1, 2, 4

(3) These values are then squared (3 x 3 = 9; 2 x 2 = 4; etc.):

9, 4, 4, 1, 0, 0, 0, 1, 1, 4, 16

(4) The sum of these squared differences (9 + 4 + ...) is 40.

(5) This sum is divided by the total number of measurements (n) minus one (n - 1 = 10):

$$\frac{40}{10} = 40$$

(6) Finally, we take the square root to obtain the standard deviation from the mean:

$$\sqrt{4} = 2 \text{ kg}$$

A large standard deviation shows that there is a wide scatter of measured values around the mean, while a small standard deviation shows that the individual values are concentrated around the mean with little variation among them.

For instance, if the weights in **Example 1** were

10, 11, 12, 14, 16, 16, 16, 18, 20, 21 and 22 kg,

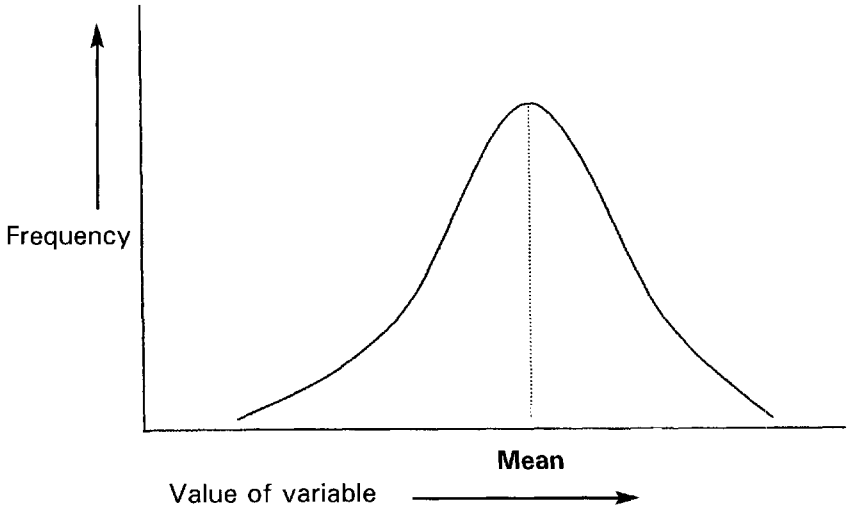
the mean weight would still have been 16 kg. However, the standard deviation would have been 4 kg, indicating a much larger variation in the observations.

The formula just used to calculate the standard deviation is presented in **Annex 27.1**. Another way to calculate a standard deviation is to use the formula that is given in **Annex 27.2**. Fortunately many pocket calculators can do this calculation for us, but it is still important to understand what it means.

III. THE NORMAL DISTRIBUTION

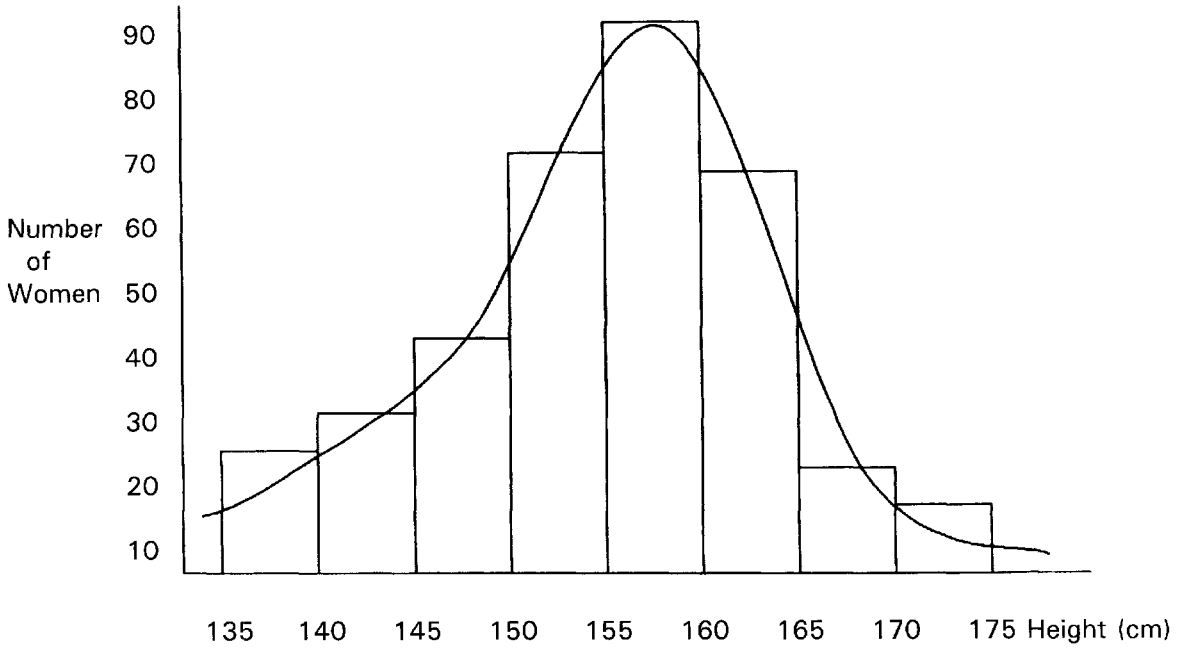
Many variables such as height, weight and age have a **normal distribution**. This distribution is a bell shaped curve with most of the values clustered near the mean and a few values out near the tails. The normal distribution is symmetrical around the mean. The mean, the median and the mode of a normal distribution have the same value.

Figure 27.1: Normal distribution curve



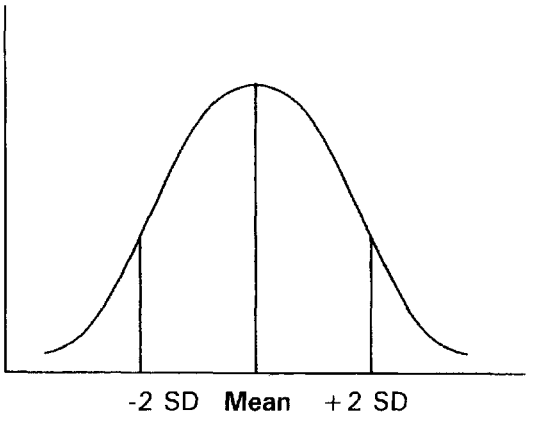
In **Figure 27.2** a histogram of the heights of pregnant women attending an antenatal clinic is shown, with the normal curve drawn over it.

Figure 27.2: Heights of pregnant women attending an antenatal clinic



An important characteristic of a normally distributed variable is that 95% of the measurements have values that are approximately within 2 standard deviations (S.D.) of the mean.* This is shown in the figure below.

Figure 27.3: A normally distributed variable



Example 2:

If the mean height of a group of 120 women is 158 cm and the standard deviation is 3 cm, it means that about 95% of the women are between 152 and 164 cm (assuming that the heights are normally distributed). In other words, 2.5% of the women (which in this case corresponds to 3 women) are shorter than 152 cm and 2.5% (or 3 women) are taller than 164 cm.

* To be more precise, 95% of the measurements have values that are within 1.96 standard deviations from the mean.

Many statistical tests require that the variables be normally distributed. Therefore it is important to examine the frequency distributions for your variables to determine which of them do not have approximately normal distributions. Out of the examples given in the modules so far there are several which are not normally distributed (e.g., **Example 4** in **Module 22**). To assess whether a variable is normally distributed one can plot it as a histogram or a line graph (see **Module 22**).

IV. THE RELATION OF THE SAMPLE TO THE WHOLE STUDY POPULATION

When you undertake a study it is usually necessary to draw a sample from your study population(s). You will then describe each population on the basis of the information collected from the sample. In other words, you will try to generalise the findings from the sample to the larger study population. Obviously, this can only be done if the sample is selected in such a way that it can be considered representative of the whole study population.

Any value of a variable obtained from the sample (e.g., a sample mean) can then be considered as an estimate of the corresponding study population value (population mean).

> **For example**, if 158 cm is the calculated mean height of a sample of 120 women in a certain town you hope it is a good approximation of the mean height of the whole population of women in that town.

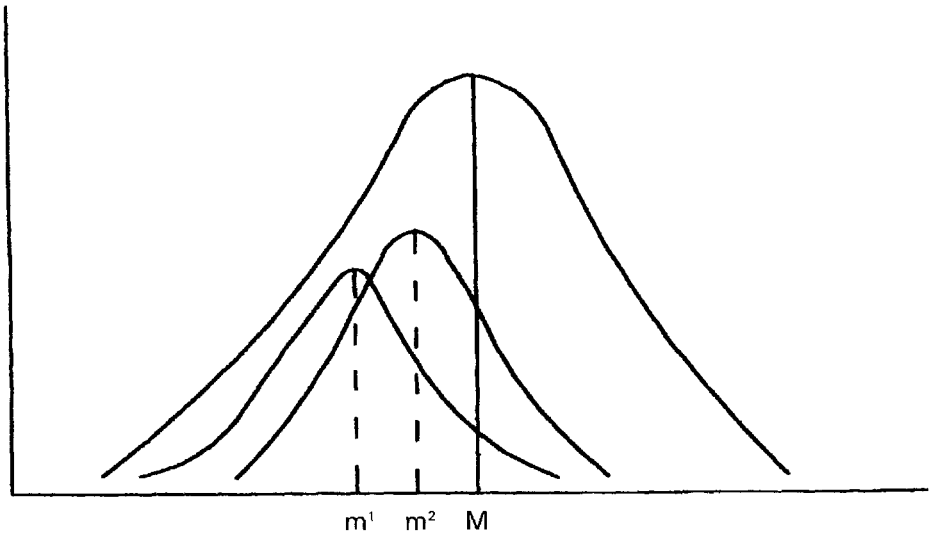
However, the sample mean is not likely to be identical to the population mean.

> If you draw another sample of 120 mothers, you might find a mean of 157 cm, which is not identical to the first sample mean. It probably also differs from the true mean height of the total population from which the sample was drawn.

This phenomenon is called **SAMPLING VARIATION.**

Figure 27.4 shows a frequency distribution curve of a population with the curves of two different samples inside it.

Figure 27.4: A frequency distribution curve and two sample curves for a variable that is normally distributed



Note that any representative sample will have a distribution curve similarly shaped to the population curve, but it can fall anywhere within the population curve.

V. HOW TO DETERMINE THE EXTENT TO WHICH THE FINDINGS IN THE SAMPLE ARE REPRESENTATIVE FOR THE STUDY POPULATION

To find out to what extent a particular sample value deviates from the population value, a range or an interval around the sample value can be worked out which will most probably contain the population value.

This range or interval is called the **CONFIDENCE INTERVAL.**

> A CONFIDENCE INTERVAL is the interval or range of values that most likely encompasses the true population value. The lower and upper limits of this interval are termed **confidence limits.**

For example:

A 95% confidence interval of 152 to 164 cm for the mean height of a population of women means that you are 95% certain that the real population mean, which you cannot know exactly unless you measure the heights of all women, lies between 152 and 164 cm. (152 cm is the lower confidence limit, 164 cm is the upper confidence limit.)

The calculation of a confidence interval takes into account the **STANDARD ERROR**. The standard error gives an estimate of the degree to which the sample mean varies from the population mean. It is computed on the basis of the standard deviation.

We will now discuss how to calculate:

- the standard error and the 95% confidence interval of a mean (for numerical data), and
- the standard error and the 95% confidence interval of a percentage (for categorical data).

1. How to calculate the standard error and 95% confidence interval of a mean

When dealing with **numerical data** you may wish to estimate to what degree the sample mean varies from the population mean.

The standard error for the mean is calculated by dividing the standard deviation by the square root of the sample size:

$$SE = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} \quad \text{or} \quad \frac{SD}{\sqrt{n}}$$

It can be assumed, for a normally distributed variable, that approximately 95% of all possible sample means lie within two standard errors of the population mean. In other words, we can be 95% sure that the population mean, of which we want to have the best possible estimate, lies within two standard errors of our sample mean.

When describing variables statistically you usually present the calculated sample mean plus or minus two standard errors. This is then called **the 95% CONFIDENCE INTERVAL**. It means that you are about 95% certain that the true population mean is within this interval.

In **Example 1** the weights of a random sample of 11 three-year-old children were taken in a village. The sample mean was 16 kg and the standard deviation of the sample was 2 kg.

The standard error is:

$$\frac{2}{\sqrt{11}} = 0.6 \text{ kg}$$

The 95% confidence interval is:

$$16 \pm (2 \times 0.6) = 14.8 \text{ to } 17.2 \text{ kg}$$

This means that we are approximately 95% certain that the mean weight of all three-year-old children in your population lies between 14.8 and 17.2 kg.

Note that the larger the sample size, the smaller the standard error and the narrower the confidence interval (or the margin of error, see **Module 11**, sample size) will be. Thus the advantage of having a sufficiently large sample is that the sample mean will be a better estimate of the population mean. However, at a certain point increases in sample size demand vast investments in time and money, whereas the confidence interval only marginally decreases. (See **Annex 11.3.**)

In the above example an increase in sample size would clearly increase the reliability of the calculation. With a sample size of 20 (instead of 11), the standard error would have been:

$$SE = \frac{2}{\sqrt{20}} = 0.45 \text{ kg},$$

and the 95% confidence interval for the mean weight would have been from 15.1 to 16.9 kg.

2. How to calculate the standard error and 95% confidence interval of a percentage

In the previous section we calculated the standard error and the 95% confidence interval of a sample mean, starting with **numerical data.** We will now do the same for a percentage that was calculated from **categorical data.**

Example 3:

Among a sample of 120 TB patients, which was drawn from the total population of TB patients in the country, it was found that 28 (or 23.3%) did not comply with their out-patient treatment. The other 92 (or 76.7%) exhibited a satisfactory degree of compliance. We now want to calculate the standard error of the percentage of non-compliance (23.3%). This is done as follows.

If p represents one of the percentages (23.3%) and 100 - p represents the other (76.7%), then the standard error of the percentage is obtained by multiplying them, dividing the result by the number in the sample and taking the square root.

The formula for the standard error of a percentage is:

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

In the **example** this is:

$$\sqrt{\frac{23.3x76.7}{120}} = 3.9$$

We now also want to calculate the confidence interval for the percentage of non-compliance in the whole country.

The 95% confidence interval is

23.3% ± (2 x 3.9) which is between 15.5% and 31.1%.

This means that we are 95% confident that in the population of all TB patients in the country from which the sample of 120 was drawn, 15.5% to 31.1% do not comply with their out-patient treatment.

Note that instead of percentages we can use proportions, which can take on any value between 0 and 1. The formula for the standard error would then be:

$$\sqrt{\frac{p(1-p)}{n}}$$

In the **example** this is:

$$SE = \sqrt{\frac{0.233x0.767}{120}} = 0.039$$

The 95% confidence interval would be between 0.155 and 0.311, or between 15.5 and 31.1%.

GROUP WORK

1. Calculate the range, the standard deviation and the 95% confidence interval for your most important sets of numerical data.

   Interpret the results of these calculations.

2. Calculate the 95% confidence interval of percentages for your most important sets of categorical data. This same calculation can be made for numerical data if they are summarised in categories.

   Interpret the results of the calculations.

3. Save the results of your group work. You will need them when you perform statistical tests.

Annex 27.1: Long method for the calculation of a standard deviation

x_i = the observations (weight of three year old children, see example 1)

x = mean = $\frac{\sum x_i}{n}$

$x_i - x$ = difference = di

The mean x = 176/11 = 16 kg

x_i	$x - x_i$	di	di^2
13	16 – 13	+3	9
14	16 – 14	+2	4
14	16 – 14	+2	4
15	16 – 15	+1	1
16	16 – 16	0	0
16	16 – 16	0	0
16	16 – 16	0	0
17	16 – 17	- 1	1
17	16 – 17	- 1	1
18	16 – 18	- 2	4
20	16 – 20	- 4	16
			40

$\sum di^2 = 40$

The standard deviation is:

$$\sqrt{\frac{\sum di^2}{n - 1}} = \sqrt{40/10} = 2 \text{ kg}$$

Annex 27.2: Short method for calculation of a standard deviation

$$\text{Standard deviation (S.D.)} = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n - 1}}$$

where x = each value
x^2 = the square of each value
(sigma) $\sum$ stands for 'the sum of' and
n = the number of observations.

If we apply this formula to **Example 1** (weights of 11 three year old children on page 5), you will find that it is not as difficult as it first seems. The values x are:

13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 20 kg and n = 11

If we square each value we get:

169, 196, 196, 225, 256, 256, 256, 289, 289, 324, 400

The sum of all the squares ($\sum x^2$) is 2856

The sum of all the values ($\sum x$) is 176

Therefore $(\sum x)^2 = 176^2 = 30{,}976$

and $\frac{(\sum x)^2}{n} = \frac{30{,}976}{11} = 2816$

$$\text{Standard deviation} = \sqrt{\frac{(2856 - 2816)}{10}} = \sqrt{4} = 2 \text{ kg}$$

Note:

A faster method to calculate the standard deviation is to use the automatic function built into many pocket calculators.

Trainer's Notes

Module 27: PREPARATION FOR STATISTICAL ANALYSIS

Timing and teaching methods

1 hour	Introduction and discussion
3 hours +	Group work

Introduction and discussion

- This is the first module in which statistical concepts are introduced which will be new for some, if not all, of the participants, depending on their academic level. Therefore it is recommended that you go slowly, be careful with formulae and use simple examples.

- Be sure that the concept of *sampling variation* (part IV) is clear to everybody since this is crucial for a good understanding of how significance tests work (**Module 28**).

- Make the link with **Module 11** and show **Annex 11.3** to visualise the diminishing returns of ever increasing sample sizes.

- If participants have little experience with statistics you may decide to leave out section 1 and 2 of Part V from your presentation. However, if one of the groups is doing a descriptive study in which no attempt is made to compare groups, they may need to calculate the standard error of a mean (section 1) or the standard error of a percentage (section 2), instead of performing statistical tests. In that case you should leave in this section when you make your presentation.

Group work

- If some of the calculations are not appropriate for certain groups they should do at least one of each calculations using sample data for the sake of experience.

- Note that there is no need to calculate standard errors of differences or percentages if statistical tests are to be performed to determine differences between groups or to measure associations between variables. (See **Modules 28 - 31**.)

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 28

CHOOSING A SIGNIFICANCE TEST

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 28: CHOOSING A SIGNIFICANCE TEST

OBJECTIVES

At the end of this session you should be able to:

1. **Explain** what a significance test is and what its purpose is.
2. **Use** the tables to choose appropriate significance tests for different sets of data.
3. **Choose** appropriate significance tests for your own data.

I. INTRODUCTION

Throughout this course it has been stressed that the analysis and interpretation of the results of our study must be related to the study objectives. **Module 24** explained how to construct cross-tabulations appropriate to the research objectives. We may have found some interesting results. **For example**, in a study on alcohol use, we find that 30% of the men included in the sample are drinking more than three glasses of alcohol a day compared to only 20% of the women. How should we interpret this result?

- The observed difference of 10% might be a **TRUE DIFFERENCE**, which also exists in the total population from which the sample was drawn.

- The difference might also be **DUE TO CHANCE**; in reality there is no difference between men and women, but the sample of men just happened to differ from the sample of women. One can also say that the observed difference is due to sampling variation.

- A third possibility is that the observed difference of 10% is due to defects in the study design (also referred to as **BIAS**). For example, we only used male interviewers, or omitted a pre-test so we did not discover that alcoholism is a very sensitive topic for women which requires a flexible approach. With an appropriate study design no such difference would have been found.

If we feel confident that an observed difference between two groups cannot be explained by bias, we would like to find out whether this difference can be considered as a true difference. We can only conclude that this is the case if we can **rule out chance** (sampling variation, see **Module 27**) as an explanation. We accomplish this by applying a significance test.

A SIGNIFICANCE TEST estimates the likelihood that an observed study result (e.g., a difference between two groups) is due to chance.

In other words, a significance test is used to find out whether a study result, which is observed in a sample can be considered as a result which indeed exists in the study population from which the sample was drawn.

II. SIGNIFICANCE TESTS

Different sets of data require different significance tests. Throughout this module and **Modules 29-31**, two major sets of data will be distinguished.

- Two (or more) **GROUPS**, which will be compared to detect **DIFFERENCES**. (E.g., men and women, compared to detect differences in alcohol used.)

- Two (or more) **VARIABLES**, which will be measured in order to detect if there is an **ASSOCIATION** between them. (E.g., between alcohol use and income.)

In order to help you choose the right test, a flowchart and matrices will be presented for different sets of data. In the modules that follow (29-31) some common significance tests will be further explained. First, however, we will discuss how significance tests work.

III. HOW SIGNIFICANCE TESTS WORK

The reasoning behind significance tests is the same, no matter whether a researcher is comparing two groups for differences or whether (s)he is measuring two variables to detect possible associations.

We will first concentrate on the **comparison of groups**.

- Suppose you observed a difference between two groups in your sample.
- You want to know whether this observed difference between the two groups represents a *real* difference in the total study population from which the sample was drawn, or whether it just *occurred by chance* (due to sampling variation).
- To find this out, you determine how likely it is that this difference could have occurred by chance, if in the total population no difference exists between the two groups.

We are never 100% sure that an observed difference is true. In general, we are happy if we can be 99% or 95% sure (confident) that the observed result is true for the whole study population. If we are 95% sure, there is a less than 5% likelihood that the observed difference occurred by chance. We usually choose a commonly accepted level of allowing that our conclusion may have occurred by chance, 0.10 (10%), 0.05 (5%), 0.01 (1%) or even 0.001 (0.1%). This is called the **chosen significance level** (also called α, or alpha level). The term **confidence level** may be used as well.

If you want to be as sure as possible that the difference you observed is true, you will choose the 0.01 or 0.001 level of significance. A trial of new drugs would choose those levels. In many management or behavioural studies we cannot be so sure, and we usually accept the 0.05 level of significance. The 0.01 and 0.05 values are most commonly used in scientific studies.

In any study looking for differences between groups or associations between variables, the likelihood or PROBABILITY (p) of observing a certain result by chance has to be calculated by statistical tests.

This PROBABILITY of observing a result by chance is usually expressed as a P-VALUE

In the **alcohol study**, the calculated p value determining whether the observed difference between men and women in their drinking behaviour was due to chance was 0.009.

The chosen significance level was 0.01 (1%), which is lower than 0.009.
We can therefore be more than 99% sure that men are drinking more heavily than women in the selected study population. We then say that **this result is statistically significant at the 0.01 level.**

If the p value had been higher than 0.01 (e.g., 0.03), the result would **not** have been statistically significant at the 0.01 level.

Note that the same reasoning applies for **associations between variables.**

Suppose we find that we have categorised drinking behaviour in four groups: never, moderate, severe, excessive. Income groups have been categorised as low, moderate and high. We observe a trend that drinking behaviour is more severe or excessive in lower income groups. The probability that this association between alcohol use and income occurs by chance will now have to be calculated. The calculated p value is 0.07. We had chosen a significance level of 0.05%. As our p value is higher than 0.05, this result is not statistically significant at the 0.05 level, and we cannot be 95% sure that the association between alcohol use and income is a real one.

Note:

In statistical terms the assumption that in the total study population **no** real difference exists between groups (or that no real association exists between variables) is called the **NULL HYPOTHESIS (Ho)**. The **ALTERNATIVE HYPOTHESIS (Ha)** is that there exists a difference between groups or that a real association exists between variables

Examples of null hypotheses are:

- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males do not drink more alcohol than females.
- There is no association between families' income and malnutrition in their children.

Note:

If the result is statistically significant, we reject the **NULL HYPOTHESIS (Ha)** and accept the **ALTERNATIVE HYPOTHESIS (Ha)** that there *is* real difference between two groups, or a real association between two variables.

Examples of alternative hypotheses are:

- There is a difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males drink more alcohol than females.
- There is an association between families' income and malnutrition in their children.

Be aware that 'statistically significant' does not mean that a difference or an association is *of practical importance.* The tiniest and most irrelevant difference will turn out to be statistically significant if a big enough sample is taken. On the other hand, a large and important difference may fail to reach statistical significance if too small a sample is used.

It is important for the researcher to state whether the results that are statistically significant are also practically significant. Practical significance implies that the problem warrants some action to be taken to alleviate it. For example, if the difference in the incidence of measles between 500 vaccinated and 500 non-vaccinated children is 35%, then vaccination ought to be promoted more actively. However, it the difference is only 2%, it may not warrant any additional action.

IV. CHOOSING A SIGNIFICANCE TEST

Depending on the aim of your study and the type of data collected, you have to choose an appropriate significance test.

Note:

Before applying any statistical test, state the null hypothesis in relation to the data to which the test is being applied. This will enable you to interpret the results of the test.

The following sections will explain how you will choose an appropriate statistical test to determine differences between groups or associations between variables.

1. Determining significant differences between groups (using **Figure 28.1**)

When deciding what test to use to determine whether differences between groups are statistically significant, there are several issues you must consider. First you need to decide whether you have paired* or unpaired observations. (See **Modules 9** and **26**, if necessary.) Within each of these categories it is necessary to determine whether the data are categorical (nominal and ordinal) or numerical. (For definitions see **Modules 8** and **22**.)

For NOMINAL data (paired or unpaired) the significance test to be used depends on whether the sample is small or large. There is no clear guide to what should be considered 'small' or 'large'. However, in the case of unpaired observations it is better to use **Fisher's exact test** rather than the **Chi-square test** if the total sample is less than 40 *or* if any cell of the table, which must be constructed, has an expected number of less than 5. In these training modules we will only consider the tests used with larger samples. The **Chi-square test** will be dealt with in **Module 29** and **McNemar's chi-square test** in **Module 30**.

Example of an unpaired sample:

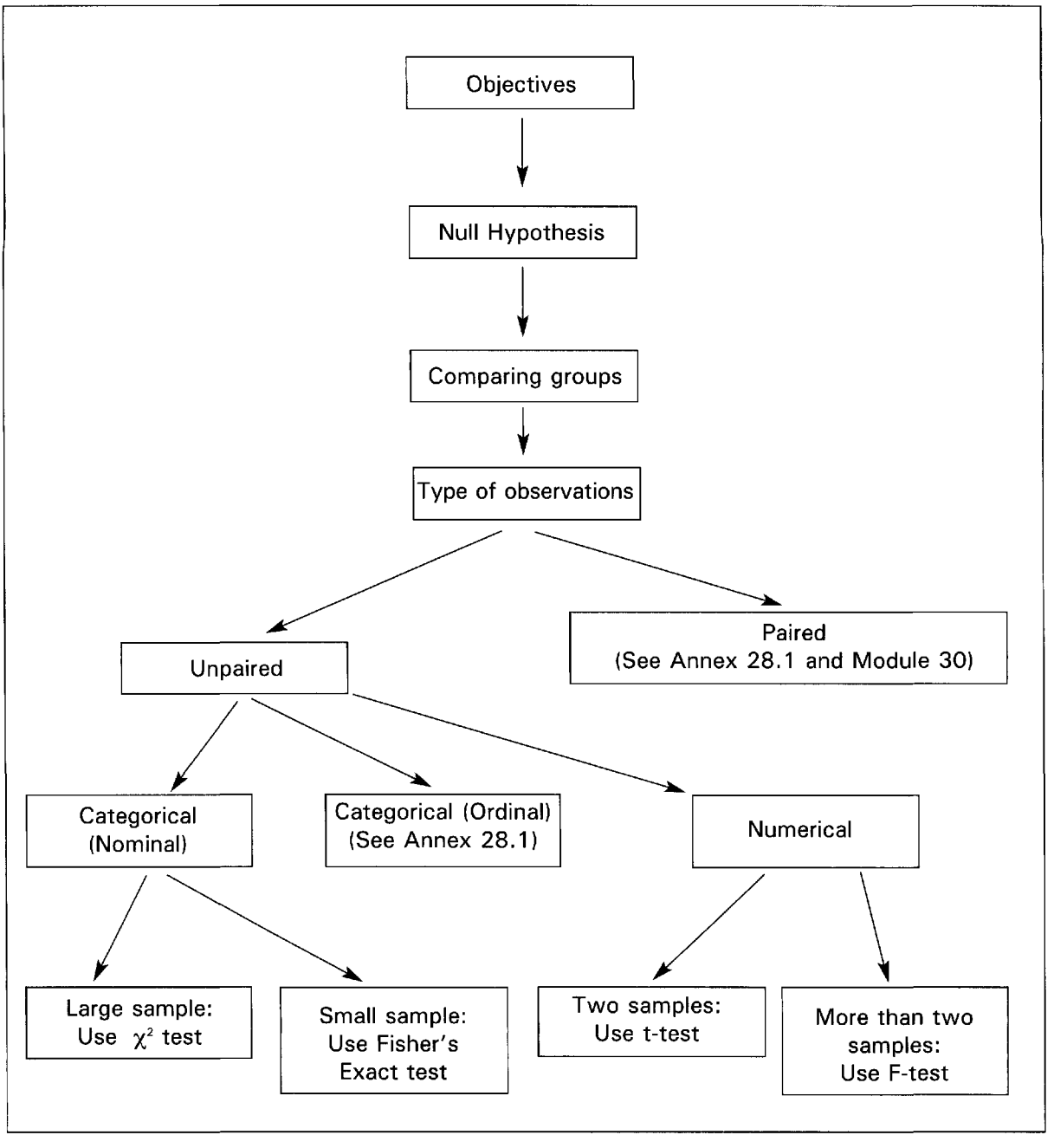
In a study of the effectiveness of measles vaccination the research team decided to study 100 measles patients aged 1-5 years coming to a clinic and 100 children in the same age range who came with other diseases. When comparing the two groups for their vaccination status, they found that the vaccination rate among children with measles was lower than among non-measles patients. The **Chi-square test** was used to test whether this difference was statistically significant.

Using **Figure 28.1**, these are the steps to follow, in determining which test to use.

1. *Objective*: To study the effectiveness of the measles vaccines among children 1 to 5 year old.
2. *Hypothesis:*
   H_0: There is no difference in the vaccination rate among the measles and the non-measles patients.
   H_a: There is a difference in the vaccination rate among the measles and the non-measles patients.
3. *Data set*: Comparing groups (vaccination rate).
4. *Type of observation*: Unpaired.
5. *Type of variable*: Categorical (Nominal).
6. *Sample*: Large.
7. *Test:* χ^2 Test

* In paired observations individual observations in one data set (e.g., cases) are matched with individual observations in another data set (e.g., controls), for example by taking care that both participants in the study come from the same location or have the same age.

Figure 28.1: Flow chart for choosing a significance test when determining differences between groups



For ORDINAL data the significance test to be used depends on whether only two groups or more than two groups are being compared.*

For NUMERICAL data, as for ordinal data, the choice of an appropriate significance test depends on whether you are comparing two groups or more than two groups.

* The tests to be used for comparing two groups are based on ranking of data: **Wilcoxon's two-sample test**, which gives equivalent results to the **Mann-Whitney U-test**, for unpaired observations and **Wilcoxon's signed rank test** for paired observations. We will not deal with these tests in our training modules, but if you want to use them, please refer to a textbook on statistics. (See References to this module; Chapter 10 of Swinscow's *Statistics at Square One* is very clear and easy to understand.)

In **Module 29** we will discuss how to conduct and interpret a **t-test** if you are comparing only two groups. If you are comparing more than two groups, you should refer to a textbook on statistics. (See **Figure 28.1** and **Table 28.1**).

Example of an unpaired sample, two groups:

In a nutrition study the weights of 142 five year-olds living in rural areas and of 171 five year-olds living in urban areas were measured. The mean weight for each of the two samples was calculated and compared, using the **t-test**, to determine whether there was a difference.

Using **Figure 28.1**, these are the steps to follow, in determining which test to use.

1. *Objective:* To determine the weight difference between five year olds in rural and urban areas.
2. *Hypothesis:*
   H_0: There is no difference between the mean weights of children in urban and rural areas.
   H_a: There is a difference between the mean weights of children in urban and rural areas.
3. *Data set:* Comparing groups (mean weight)
4. *Type of observation:* Unpaired.
5. *Type of variable:* Numerical.
6. *Sample:* Two samples.
7. *Test:* t- test.

Example of an unpaired sample, more than two groups:

The mean weight of the following 4 groups of five-year-olds were compared: boys living in rural areas, boys living in urban areas, girls living in rural areas and girls living in urban areas. In this case the **F-test** was the appropriate choice.

Example of a study in which data are ranked:

In a quasi-experimental study to investigate the effect of a health education campaign on the knowledge of management of diarrhoea in the community, two groups of villages were selected. The first group was composed of villages in which the campaign was held, the second group of villages in which no health education was given. During analysis the villages were ranked from the highest level of knowledge of adequate treatment of diarrhoea to the lowest. The **Wilcoxon's two-sample test** was performed to determine whether there was a significant difference between the two groups of villages.

Example of a paired sample, two groups:

The mean weights of adult males and adult females were compared while controlling for height. This meant that for each male of a certain height a female of the same height was selected so that each pair could be compared on weight. The **paired t-test** was used in this instance.

Table 28.1 summarises the different tests available when determining whether the differences between two or more groups with paired or unpaired observations are significant.

Table 28.1: Choosing a significance test when determining differences between groups

	Unpaired observations	Paired observations
Nominal data		
Small sample	Fisher's exact test	Sign test
Large sample	Chi-square test* (Module 29)	McNemar's chi-square test* (Module 30)
Ordinal data		
Two groups	Wilcoxon two-sample test or Mann-Whitney U-test	Wilcoxon sign-rank test
More than two groups	Kruskal-Wallis 1-way analysis of variance	Friedman 2-way analysis of variance
Numerical data		
Two groups	t-test* (Module 29)	Paired t-test* (Module 30)
More than two groups	F-test	

* Tests indicated by asterisks will be discussed in the modules indicated.
Source: Adapted from Riegelman RK (1981).

2. Measuring associations between variables (using Table 28.2)

Table 28.2: Choosing a significance test when measuring associations between variables.

Nominal data	Chi-square test (if sample is large enough)* (Modules 29, 30)	Calculate odds ratio or estimate relative risk* (Module 25)
Ordinal data or *numerical data* when *no linear relationship* is suspected	Calculate Spearman's rho or Kendall's tau	Significance of Spearman's rho or Kendall's tau
Numerical data when a *linear relationship* is suspected	Calculate Pearson's correlation coefficient (r)* (Module 31)	Significance of Pearson's correlation coefficient (r)* (Module 31)

* Tests indicated with an asterix will be discussed in the modules indicated.
Source: Adapted from Riegelman RK (1981).

Determine whether your data are nominal, ordinal or numerical. If they are numerical, decide whether a linear relationship is suspected. The term 'linear relationship' for numerical data means that the association is such that the dependent variable changes in a constant relationship to the independent variable in such a way that the points on a scatter diagram, when joined, are approximated best by a straight line.

For NOMINAL data the **relative risk** is a useful measure of association that is often applied in case-control and cohort studies. **Module 25** dealt with calculating the **odds ratio** as an estimate of relative risk in case-control studies.

Example:

In the case-control study presented in **Example 3** of that module, we used the **odds ratio** to calculate that neonates with a low birth would have a 12 times higher risk of neonatal death than babies with a normal weight.

For ORDINAL data Spearman's rank correlation coefficient (rho) or **Kendall's tau** can be calculated and tested for significance. If you want to use them, refer to statistic textbooks. (See the references, e.g. Swinscow (1998) Chapter 11.)

For NUMERICAL data when a linear relationship is suspected **Pearson's correlation coefficient** can be calculated and tested for significance. **Module 31** will discuss how to do this.

Example:

You may want to examine whether the weights of five-year old children are associated with their families' income. You suspect that there is a linear relationship between the two variables 'family income' and 'weight', such that weight increases with increasing family income.

Note:

An association that is statistically significant does not necessarily imply the existence of a causal relationship. (See **Module 25**.) However, it often invites further investigation to find out whether or not a causal relationship does exist.

EXERCISES on choosing tests

Using **Figure 28.1**, and **Tables 28.1** and **28.2**, identify the appropriate tests for the following research studies.

Exercise 1:

A study will be undertaken to compare the effect of a new anti-hypertensive drug on the diastolic blood pressure of a study group sample compared to the effect of a placebo on an unmatched control group sample (Riegelman 1981: 243).

Exercise 2:

A study will be conducted to find out whether pregnant women living in households where there is no water supply of their own are at significantly greater risk of experiencing perinatal deaths than pregnant women who live in households with an independent water supply. if so, the study will measure how strong this association is.

Exercise 3:

A study was undertaken to determine whether there was a significant weight loss after a one year course of therapy for diabetes, and whether the amount of weight loss was related to initial weight. The following table gives the initial weights (x) and weights after one year of therapy (y) for 16 newly diagnosed adult diabetic patients.

Initial weight (x) in pounds	Weight after 1 year (y) in pounds	Initial weight (x) in pounds	Weight after 1 year (y) in pounds
140	115	120	123
160	130	145	143
180	135	150	125
120	125	160	140
132	112	160	135
146	130	149	120
190	160	129	119
200	160	150	113

Exercise 4:

From previous studies it is determined that 30% of the eligible couples in a Health District practise family planning. After a mass educational programme, results indicated that out of 90 eligible couples randomly selected, 40 practised family planning. The Health Education Officer wishes to know whether his programme has had an impact on the target group.

GROUP WORK

Referring to the specific objectives of your study and the list of variables and using the cross-tabulations already made, identify the significance tests you will need to perform on your data.

REFERENCES

1. Altman DG (1991) *Practical statistics for medical research*. London: Chapman and Hall.

2. Barker DJP (1982) *Practical epidemiology*. (3rd ed.). Edinburgh, UK: Churchill Livingstone.

3. Castle WM and North PM (1995) *Statistics in Small Doses.* Edinburgh, UK: Churchill Livingstone.

4. Bradford Hill A (1984) *A short textbook of medical statistics* (11th ed.). London, UK: Hodder and Stoughton.

5. Kelsey JL, Thompson WD and Evans AS (1986) *Methods in observational epidemiology*. Oxford, UK: Oxford University Press.

6. Kidder LH and Judd CM (1986) *Research methods in social relations*. New York, USA: CBS College Publishing.

7. Kleinbaum DG, Kupper LL, Morgenstern H (1982) *Epidemiologic research - principles and quantitative methods*. New York, USA: Van Nostrald Reinhold.

8. Riegelman RF (1981) *Studying a study and testing a test*. Boston, MA, USA: Little Brown and Company.

9. Schlesselman JJ (1982) *Case-control studies - design, conduct, analysis*. Oxford, UK: Oxford University Press.

10. Swinscow TDV, revised by MJ Campbell (1998) *Statistics at square one* (11th ed.). London, UK: British Medical Association.

Trainer's Notes

Module 28: CHOOSING A SIGNIFICANCE TEST

Timing and teaching methods

½ hour* Introduction and discussion
1 hour* Group work

* ½ to 1 hour should be added to either the plenary or the group work if the trainer decides to ask the participants to complete the exercises on choosing a significance test in one of these two components of the session.

Introduction and discussion

- This module should not necessarily be presented in full. As stated in the objectives, the main aim of the session is that participants understand **what** significance tests are and **why** they would use them. They should be able to use the flowcharts to choose the appropriate significance tests for different study designs and different types of data. The examples given for the use of the two flow charts should **not** be all presented.

- The question of **why** you perform significance tests can be introduced by presenting a cross-table (for example the numbers of smokers and non-smokers among males and females) and subsequent asking how the difference between males and females (30% versus 20%) can be interpreted. It would even be better to take a cross-table of one of the research teams themselves and ask the same question.

- Once it is clear to everybody why significance tests are performed, you might ask participants to give examples of results (cross-tables) from their own projects for which significance tests have to be performed.

- Stress that there are two tests that are most commonly used: the t-test and the χ^2 test. All other tests mentioned in the flowchart (**Figure 28.1**) are less likely to be used in the types of projects the participants most often develop.

- If your group of participants are advanced enough to learn to use the flowchart and the matrices on their own, you could ask them to complete several or all of the five exercises. They could be asked to take a few minutes during plenary to use the flowchart and the matrices to select the appropriate tests, and then volunteers could be asked for the answers. Alternatively, several or all of the exercises could be the first task during group work, with the facilitator playing an active role in assisting group members in becoming adept at using the flow chart and tables.

- After having introduced this module we *recommend that you present part of the next module: either the t-test or the chi-square test. This will make the theoretical concepts discussed in this module more concrete.*

Group work

Let the participants decide for which cross-tabulations they should perform significance tests. They should determine which test is appropriate for each cross-tabulation selected.

Suggested tests for research studies given in exercises

Exercise 1:

The study deals with two **samples**. We are interested in significant differences in diastolic blood pressure between the study group that got the new drug and the comparison group that got the placebo. This is **numerical data** (**Table 28.1** or **Figure 28.1**).

The samples are **unmatched**. Hence the **t-test** is the appropriate test.

Exercise 2:

We assume that these are two **samples**. It is **differences** that are being studied for significance, and the outcome or dependent variable studied is number of perinatal deaths, i.e., **nominal** data (**Figure 28.1** or **Table 28.1**). Such studies are done on large samples and these are **unmatched**. Hence the test is **chi-square**.

If we want to find the strength of **association** we should use Table 28.2. As this is **nominal data** the **odds ratio** or **relative risk** must be calculated. The same 2 x 2 table constructed for χ^2 test can be used here also.

Exercise 3:

There are two samples and the differences are being examined for statistical significance. The dependent variable is weight, which is numerical. (Use Figure 28.1 or Table 28.1.) There was only one sample of 16 patients, but measurements were done twice on each patient. This is thus a **matched (or paired) sample**. In this case it was self-pairing.

Hence the test is **matched t-test**. In regards to whether the weight loss was related to initial weight, it is a test of degree of **association**.

Because the data are **numerical data**, calculate **Pearson's correlation coefficient (r)**. If it is necessary to test statistical significance of association, use **Table 28.2**, which leads to statistical significance of **Pearson's r**.

Exercise 4:

In this study one **sample** has been selected. It is compared with population data that is known. The test should deal with differences between proportions or percentages. Because the data are **nominal**, the χ^2 test will be used. (See **Figure 28.1** or **Table 28.1**.)

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 29

DETERMINING DIFFERENCES BETWEEN GROUPS: PART I ANALYSIS OF UNPAIRED OBSERVATIONS

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 29: DETERMINING DIFFERENCES BETWEEN GROUPS, PART I: ANALYSIS OF UNPAIRED OBSERVATIONS

OBJECTIVES:

At the end of this session you should be able to:

1. **Decide** when to apply the unpaired t-test and the chi-square test.
2. **Calculate** t-values and chi-square values.
3. **Use** the t-tables and chi-square tables to assess whether calculated t- and chi-square values are significant.
4. **Make a decision** concerning whether you can use these tests on your data and, if so, what test should be used on which data.
5. **Perform** these tests on your data.

I. Introduction

II. T-test

III. Chi-square (χ^2) test

I. INTRODUCTION

When describing and analysing the cross-tabulations of your major variables (see **Modules 22 and 24**), you probably have observed differences between groups. You may want to find out if these differences are likely to be due to chance, or if they are real (statistically significant) differences.

In order to determine this, you can perform two types of tests. These are:

- the t-test, and
- the chi-square (χ^2) test

The **t-test** is used for NUMERICAL data, when comparing the means of two groups.

The **chi-square test** is used for CATEGORICAL data, when comparing proportions of events occurring in two or more groups.

Both tests are used for UNPAIRED observations. For observations that are PAIRED, two different tests are used, again depending on whether the data is categorical or numerical. (See **Module 30**.)

II. T-TEST

The **t-test**, also referred to as **Student's t-test**, is used for numerical data to determine whether an observed difference between the means of two groups can be considered statistically significant.

Example 1:

It has been observed that in a certain province the proportion of women who are delivered through Cesarean section is very high. A study is therefore conducted to discover why this is the case. As small height is known to be one of the risk factors related to difficult deliveries, the researcher may want to find out if there is a difference between the mean height of women in this province who had normal deliveries and of those who had Cesarean sections. The null hypothesis would be that there is no difference between the mean heights of the two groups of women. Suppose the following results were found:

Table 29.1: Mean heights of women with normal deliveries and of women with Cesarean sections

Type of delivery	Number of women included in study	Mean height in cm	Standard deviation
Normal delivery	60	156	3.1
Cesarean section	52	154	2.8

A t-test would be the appropriate way to determine whether the observed difference of 2 cm can be considered statistically significant.

To actually perform a t-test you have to complete 3 steps:

1. Calculate the t-value
2. Choose the level of significance and use a t-table
3. Interpret the results

Step 1. Calculating the t-value

To calculate the t-value you need to complete the following tasks:

(1) **Calculate the difference between the means.**
In the above example the difference is 156-154 = 2 cm.

(2) **Calculate the standard deviation** for each of the study groups. (The concept of standard deviation and how it is calculated has already been discussed in **Module 27**). Suppose the standard deviations shown in the final column of **Table 29.1** were found.

(3) **Calculate the standard error of the difference between the two means.**

The standard error of the difference is given by the following formula:

$$\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

Where: SD_1 is the standard deviation of the first sample
SD_2 is the standard deviation of the second sample
n_1 is the sample size of the first sample
n_2 is the sample size of the second sample

For our data if we take the women with normal deliveries as sample 1 and those with Cesarean sections as sample 2 the standard error of the difference is:

$$\sqrt{\frac{3.1^2}{60} + \frac{2.8^2}{52}} = 0.56$$

(4) Finally, **divide the difference between the means by the standard error** of the difference. The value now obtained is called t-value.

In the above example: $t = \frac{2}{0.56} = 3.6$

Expressed in one single formula:

$$t = \frac{\bar{X}_1 \quad \bar{X}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

(where $\bar{X}_1$ is the mean value of the first sample, and $\bar{X}_2$ is the mean value of the second sample)

Step 2. Using a t-table

Once the t-value has been calculated, you will have to refer to a t-table, from which you can determine whether the null hypothesis is rejected or not. **Annex 29.1** contains a t-table.

(1) First, decide which **significance level** (**α-value** or alpha value) you want to use (see **Module 28**). Remember that the chosen significance level (α-value) is an expression of the likelihood of finding a difference by chance when there is no real difference. Usually we choose a significance level of 0.05.

(2) Second, determine the number of **degrees of freedom** for the test being performed. Degrees of freedom is a measure derived from the sample size, which has to be taken into account when performing a t-test. The bigger the sample size (and the degrees of freedom) the smaller the difference needed in order to reject the null hypothesis.

(3) Third, the t-value belonging to the α-value (the significance level we chose) and the degrees of freedom are located in the table.

If the calculated t-value is **equal to or larger** than the value derived from the table, the p-value of significance is **smaller** than the chosen α-value (indicated at the top of the column). We then **reject** the **null hypothesis** and conclude that there is a statistically significant difference between the two means.

If the calculated t-value is **smaller** than the value derived from the table, the p-value is **larger** than the α-level we chose. We then **accept** the **null hypothesis** and conclude that the observed difference is not statistically significant.

The way the number of degrees of freedom is calculated differs from one statistical test to the other. For **student's t-test** the number of degrees of freedom is calculated as the sum of the two sample sizes minus 2.

Thus, for **Example 1**, comparing the heights of women with and without Cesarean sections, the number of degrees of freedom is:

d.f. = 60 + 52 - 2 = 110.

Note:

This is an approximate way of determining degrees of freedom. For the exact method, refer to a statistics textbook

In our example we look up the t-value belonging to $\alpha = 0.05$ and d.f. = 120 and we find it is 1.98.

Step 3. Interpreting the result

We now compare the absolute value of the t-value calculated in Step 1 (i.e., the t-value, ignoring the sign) with the t-value derived from the table in Step 2.

In our example the t-value calculated in step 1 is 3.6, which is larger than the t-value derived from the table in step 2 (1.98). Thus the p-value is smaller than 0.05, and we therefore reject the null hypothesis and conclude that the observed difference of 2 cm between the mean heights of women with normal deliveries and women with Cesarean sections is a statistically significant difference.

We can express this conclusion in different ways:

- We can say that the probability that the observed difference of 2 cm. of height between the two groups of women is due to chance is less than 5%.

- We can also say that the difference between the two groups is 3.6 times the standard error.

If you want to compare mean values of more than two groups (e.g., heights of urban, semi-urban and rural women) you cannot use **student's t-test**. In this case you must use the **F-test**, which is not described here.

III. CHI-SQUARE (χ^2) TEST

If you have categorical data the chi-square test is used to find out whether observed differences between proportions of events in two or more groups may be considered statistically significant.

Example 2:

Suppose that in a cross-sectional study of the factors affecting the utilisation of antenatal clinics you found that 64% of the women who lived within 10 kilometres of the clinic came for antenatal care, compared to only 47% of those who lived more than 10 kilometres away. This suggests that antenatal care (ANC) is used more often by women who live close to the clinics. The complete results are presented in **Table 29.2**:

Table 29.2: Utilisation of antenatal clinics by women living far from and near the clinic

Distance from ANC	Used ANC	Did not use ANC	TOTAL
Less than 10 km	51 (64%)	29 (36%)	80 (100%)
10 km or more	35 (47%)	40 (53%)	75 (100%)
Total	86	69	155

From the table we conclude that there seems to be a difference in the use of antenatal care between those who live close to and those who live far from the clinic (64% versus 47%). We now want to know if this observed difference is statistically significant or not.

The chi-square test can be used to give us the answer. This test is based on measuring the difference between the observed frequencies and the expected frequencies **if the null hypothesis (i.e., the hypothesis of no difference) were true**.

To perform a χ^2 test you need to complete the following 3 steps:

1. calculate the χ^2 value,
2. use a χ^2 table, and
3. interpret the χ^2

Step 1. Calculating the χ^2 value

(1) Calculate the expected frequency (E) for each cell.

To find the *expected frequency E* of a cell you *multiply the row total by the column total and divide by the grand (overall) total:*

$$E = \frac{\text{row total x column total}}{\text{grand (overall) total}}$$

(2) For each cell, subtract the expected frequency from the observed frequency (O - E).

(3) For each cell, square the result of (O - E) and divide by the expected frequency E.

(4) Add the squared results calculated in step (c) for all the cells.

The formula for calculating a chi-square value (steps (b) to (d)) is as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where: O is the observed frequency (indicated in the table)
E is the expected frequency (to be calculated), and
$\sum$ (the sum of) directs you to add together the values of $(O - E)^2 / E$ for all the cells of the table.

For a two-by two table (which contains 4 cells) the formula is:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

Step 2. Using a χ^2 table

As for the t-test, the calculated χ^2 value has to be compared with a theoretical χ^2 value in order to determine whether the null hypothesis is rejected or not. **Annex 29.2** contains a table of theoretical χ^2 values.

(1) First you must decide what significance level you want to use (alpha or α-value). We usually take 0.05.

(2) Then the degrees of freedom have to be calculated. With the χ^2 test the number of degrees of freedom is related to the number of cells, i.e. the number of groups you are comparing. The number of degrees of freedom is found by multiplying the number of rows (r) minus 1 by the number of columns (c) minus 1:

d.f. = (r-1) x (c-1)

For a simple two-by-two table the number of degrees of freedom is 1:
d.f. = (2-1) x (2-1) = 1

(3) Then the χ^2 value belonging to the α-value and the number of degrees of freedom are located in the table. If the calculated χ^2 value is equal to or larger than the χ^2 value from the table then the p-value is smaller than the chose significance (α)-value. In this case, we reject the null hypothesis and conclude that there is a statistically significant difference between the groups. If the calculated χ^2 value is smaller than the χ^2 value from the table, then the p-value found is larger than the chosen significance level of 0.05. In this case, we accept the null hypothesis and conclude that the observed difference is not statistically significant.

Step 3. Applying the χ^2

As for the t-test, the null hypothesis is rejected if $p<0.05$, which is the case if the calculated χ^2 value is larger than the theoretical χ^2 value in the table.

Let us now apply the χ^2 test to the data given in **Example 2** (utilisation of antenatal care). This gives the following result:

Step 1: Calculating the χ^2 value

First the **expected frequencies** for each cell are calculated as follows:

$E_1 = 86 \times 80/155 = 44.4$ $\quad$ $E_2 = 69 \times 80/155 = 35.6$

$E_3 = 86 \times 75/155 = 41.6$ $\quad$ $E_4 = 69 \times 75/155 = 33.4$

For convenience sake the observed and expected frequencies are shown in the following table:

Table 29.3: Utilisation of antenatal clinics observed and expected frequencies

Distance from ANC	Used ANC	Did not use ANC	TOTAL
Less than 10km	$O_1 = 51$, $E_1 = 44.4$	$O_2 = 29$, $E_2 = 35.6$	80
10km or more	$O_3 = 35$, $E_3 = 41.6$	$O_4 = 40$, $E_4 = 33.4$	75
Total	86	69	155

Note that the expected frequencies refer to the values we would have expected, given the total numbers of 80 and 75 women in the two groups, **if the null hypothesis**, stating that there is **no** difference between the two groups, **were true**.

Now **the χ^2 value** can be calculated:

$$\chi^2 = \frac{(51-44.4)^2}{44.4} + \frac{(29-35.6)^2}{35.6} + \frac{(35-41.6)^2}{41.6} + \frac{(40-33.4)^2}{33.4}$$

$$= 0.98 + 1.22 + 1.05 + 1.30 = 4.55$$

Step 2: Using the χ^2 table

As we have a simple two-by-two table, the number of degrees of freedom (d.f.) is 1.

Use the table of chi-square values in **Annex 29.2**. We have decided beforehand on a level of significance of 5% (α-value = 0.05).

As the number of d.f. is 1, we look along that row in the column where p = 0.05. This gives us the value of 3.84. Our value of 4.55 is **larger** than 3.84, which means that the p value is **smaller** than 0.05.

Step 3: Interpreting the result

We can now conclude that *the women living within a distance of 10 km from the clinic utilise antenatal care significantly more often than the women living more than 10 km away.*

It is important to present your data clearly and to carefully formulate any conclusions based on statistical tests in the final report of your study.

For the above example, you could present **Table 29.2** in the report and state your conclusions in the following way:

'Table 2 indicates that 64% of the wom living within a distance of 10 km from the clinic used ante-natal care during pregnancy, compared to only 47% of women living 10 km or further away from the nearest clinic. This difference is statistically significant (χ^2 = 4.55; $p<0.05$).'

Note:

- The χ^2 test can only be applied if the sample is large enough. The general rule is that the total sample should be at least 40 and the **expected** frequencies in each of the cells should be at least 5. If this is not the case Fisher's exact test should be used. (See Chapter 9 of Swinscow's *Statistics at Square One* referenced in this module.) If the table is more than a two-by-two table, the expected frequency of 1 in 5 cells is allowed to be less than 5.

- Unlike the t-test, the χ^2 test can also be used to compare more than two groups. In that case a table with three or more rows/columns would be designed, rather than a two-by-two table.

In the above example one could decide to distinguish between three different distances: less than 5 km, 5 to 10 km and more than 10 km. The data would then be put in a two-by-three table. The number of degrees of freedom would be (3-1) x (2-1) = 2.

Quick formula

For two-by-two tables there is a quick method for calculating the Chi-square value, which can replace step 1 described above.

If the various numbers in the cross-table are represented by the following letters:

	Condition		TOTAL
	+	-	
Exposure			
Yes	a	b	E
No	c	d	F
Total	G	H	N

Where, E = a + b; F = c + d; G = a + c; H = b + d,

The quick formula for calculating the Chi-square value for a two-by-two table is:

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{N(ad-bc)^2}{EFGH}$$

Note:

Computers are helpful when dealing with large data sets. A variety of software programmes provide statistical tests, including p-values. A *statistical calculator* can also calculate the chi-squares for you.

GROUP WORK

If your data was collected by unpaired observations, identify the appropriate significance test and perform the necessary analysis.

Annex 29.1: Student's t-distribution

The first column lists the number of degrees of freedom. The headings of the other columns give the α-values for t to exceed the entry value. Use symmetry for negative values.

Degrees of freedom	t-value if chosen p (α) = 0.05	t-value if chosen p (α) = 0.01
1	12.71	63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90
18	2.10	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
30	2.04	2.76
40	2.02	2.70
60	2.00	2.66
120	1.98	2.62
Infinite	1.96	2.58

If the calculated t-value (ignoring the sign) is **larger** than the value indicated in the table, the p-value in your calculation is **smaller** than the chosen p (α-) value indicated at the top of the column.

In that case, the null hypothesis, stating that there is no difference, is **rejected**, and it can be concluded that there **is** a significant difference in the result of your study.

Annex 29.2: Table of χ^2 values

Degrees of freedom	χ^2 value if $\alpha = 0.05$	χ^2 value if $\alpha = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22

If the calculated χ^2 value is **larger** than the value indicated in the table, the p-value is **smaller** than the chosen level of significance (indicated at the top of the column).

In that case, the null hypothesis, stating that there is no difference, is **rejected**, and it can be concluded that the difference between the two groups in your study **is** statistically significant.

Annex 29.3: Dealing with confounding variables: Mantel-Haenszel Chi-Square test

In **Table 29.4** the results of a schistosomiasis survey among the inhabitants of two villages are presented.

Table 29.4: Prevalence of schistosomiasis in two villages, A and B

	Village A	Village B	TOTAL
Schisto (+)	80 (32%)	80 (32%)	160
No schisto (-)	170 (68%)	170 (68%)	340
Total	250 (100%)	250 (100%)	500

It seems that the prevalence of schistosomiasis is the same in both villages (32%).

However, the researchers suspect that age is a confounding variable. Therefore, **Table 29.4** is split up into two tables (**27.5** and **27.6**). Note that adding the numbers in Tables 29.5 and 29.6 will give us Table 29.4.

Table 29.5: Prevalence of schistosomiasis in children aged 5-19 in villages A and B

	Village A	Village B	TOTAL
Schisto (+)	37 (62%)	73 (38%)	110
No Schisto (-)	23 (38%)	117 (62%)	140
Total	60 (100%)	190 (100%)	250

χ^2 - 9.08; 1 degree of freedom; $p < 0.01$.

Table 29.6: Prevalence of schistosomiasis in those aged 20 years and above in villages A and B

	Village A	Village B	TOTAL
Schisto (+)	43 (23%)	7 (12%)	50
No schisto (-)	147 (77%)	53 (88%)	200
Total	190 (100%)	60 (100%)	250

χ^2 - 2.78; 1 degree of freedom; $p < 0.05$.

From **Tables 29.5** and **29.6** it becomes clear that:

- Within each age group schistisomiasis is more prevalent in Village A than in Village B.
- Schistosomiasis is more prevalent in children than in adults.
- In Village A there are relatively few children and many adults compared to Village B.

Age is said to be a confounding variable because it is related to the variable of interest (prevalence of schistosomiasis) and to the groups being compared (residence in Village A or B).

This example illustrates an important point in analysing data. It may be very misleading to pool dissimilar data. In this particular example, pooling the age groups masked an important real difference. In other situations pooling the data may suggest a difference or association that does not exist or even a difference opposite to that which exists.

It is, therefore, important to analyse the above data for the different age groups separately. The appropriate χ^2 values (with continuity correction) for comparing the prevalences in Villages A and B are shown in **Tables 29.5 and 29.6.** The difference in prevalence is significant for children but not for adults.

Mantel-Haenszel χ^2 test

It is often useful to have a summary test that pools the evidence from the individual tables, but takes into account the confounding factor (age in our example). The Mantel-Haenszel χ^2 test for doing this will be described.

For each of the two-by-two tables we will use the notation:

Exposure	Present (+)	Absent (-)	TOTAL
Yes (+)	a	b	E
No (-)	c	d	F
Total	G	H	N

Step 1. For each of the two-by-two tables,

1. Find the observed frequency O_a
2. Calculate the expected frequency E_a, which equals EG/N
3. Calculate the variance V_a, which equals EFGH/(N^2 (N-1))

Step 2. The Mantel-Haenszel Chi-Square (χ^2_{MH}) value is

$$\chi^2_{MH} = \frac{(O - E - 0.5)^2}{V_a} \quad \text{wiht degrees of freedom} = 1$$

Where: O = the sum of the (O_a) observed frequencies
E = the sum of the (E_a) expected frequencies
V = the sum of the (V_a) variances
0.5 is the continuity correction factor

To check for statistical significance, we use the χ^2 tables as discussed earlier.

Application:

In the prevalence of schistosomiasis in the two villages, there are two 2-by-2 tables, for those less than 20 years (children), and for those aged 20 years and above (adults). (See **Tables 29.5 and 29.6.**) From the two tables, the observed and expected frequencies are given below.

In the **example,** the calculations are:

	O_a	$E_a = EG/N$	$V_a = EFGH/(N^2(N-1))$
Children	37	26.4	$110x140x60x190/250^2x249 = 11.3$
Adults	43	38	$50x200x190x60/250^2x249 = 7.3$
Total	80	64.4	18.6

O = 80, E = 64.4, V = 18.6

$$\chi^2 = \frac{(80 - 64.4 - 0.5)^2}{18.6} = \frac{15.1^2}{18.6} = 12.25 \ (p<0.001)$$

It, therefore, can be concluded that the prevalence of schistosomiasis is significantly different in Village A and B. (Remember that this seemed not be the case when we looked at **Table 29.4,** in which the data for both adults and children were pooled.)

Validity of Mantel-Haenszel χ^2 test

The Mantel-Haenszel χ^2 test is an approximate test. The rule for assessing its adequacy is more complicated than that for the ordinary χ^2 test. Two additional values are calculated for each table and summed over the tables. These are:

1. min(E,G), that is the smaller E and G
2. max(O, G - F), that is O if G is smaller than or equal to F and F - F if G is larger.

Both these sums should differ from the total of the expected values, E_a, by at least 5. The details of the calculation for the above example are:

	min (E,G)	**max (O, G – F)**
Children	60	0
Adults	50	0
Total	110	0

These sums are 110 and 0, both of which differ from 64.4 (E_a) by more than 5. The use of the Mantel-Haenszel test is therefore valid.

Trainer's Notes

Module 29: DETERMINING DIFFERENCES BETWEEN GROUPS, PART I: ANALYSIS OF UNPAIRED OBSERVATIONS

Timing and teaching methods

1 hour	Introduction and discussion
3 hours +	Group work

Introduction and discussion

- We advise that you *present either the t-test or the χ^2 test immediately after completing **Module 28**, so that participants get a better idea of what a significance test is and how it is used.* Probably all research teams will be applying χ^2 tests on their data, but not all teams will be using the t-test. Therefore, you might present the χ^2 test in combination with **Module 28** and the t-test in another session.

- Proceed slowly, step-by-step, when explaining how significance tests work, so as not to frighten participants who have little experience. Stress that it is not important to understand WHY the calculations of t-values and χ^2-values are performed this way (actually there is not always a clear rationale, for example, for the concept of degrees of freedom). It is enough to know HOW they are done. Be careful with formulae: present them only after the step-wise calculations have been explained.

- You may use examples taken from the groups' own studies rather than Examples 1 and 2 presented in the module. Remember, however, to use simple examples, i.e., two-by-two tables and small numbers, so that they are easier to follow.

- Take extra care to explain how to use the t-table and χ^2 table and how to interpret the results. Let participants struggle themselves with the examples before providing the correct answers.

- Also pay attention to the appropriate phrasing of conclusions based on significance tests, both in cases where the results are significant and where they are not significant.

- **Annex 29.3** should not be presented unless the group is very advanced. The teams can read it and use the test, if necessary.

Group work

When performing statistical tests make sure that each member of the group does at least one test on his or her own.

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 30

DETERMINING DIFFERENCES BETWEEN GROUPS: PART II

ANALYSIS OF PAIRED OBSERVATIONS*

* Most of this module stems from the MSc course materials of the London School of Hygiene and Tropical Medicine, with permission of Richard Hayes, Betty Kirkwood and Tom Marshall.

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	**Analyse unpaired and paired observations**	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 30: DETERMINING DIFFERENCES BETWEEN GROUPS, PART II: ANALYSIS OF PAIRED OBSERVATIONS

OBJECTIVES:

At the end of this session you should be able to:

1. **Identify** research studies where pairing or matching of subjects is necessary.

2. **Identify and use** the significance tests appropriate for studies using paired data.

I. Introduction

II. Paired t-test

III. McNemar's chi-square test

I. INTRODUCTION

This module describes the most commonly used tests for paired observations pertaining to numerical and to nominal data:

- the **paired t-test** for numerical data and
- **McNemar's chi-square test** for nominal data.

You will remember from **Module 9** and **Module 26** part III, that paired or matched observations are carried out if researchers want to ensure through their study design that the relationship between two variables they are interested in is not confounded by another variable. They therefore have to sample their cases and controls in such a way that these are similar with respect to one or more potentially confounding variables.

The concept of pairing or matching subjects is illustrated by the following examples:

Example 1:

A researcher wanted to find out whether a class of students taught with audio-visual aids (AV) receive on average better grades than those who are taught without audio-visual aids. To minimise the effect of confounding variables such as *social status* and *previous knowledge of the subjects,* each student in the AV class was paired with another in the non-AV class of similar social status and knowledge level.

Example 2:

During a nutritional survey, a quality control exercise was carried out to check the agreement between two observers in measuring the children's weight. In this instance we have paired observations as we have a *set of two observations on the same child.*

Example 3:

A study team compared schistosomiasis egg counts in two villages. It recognised that egg counts vary with age and sex. It decided to ensure that the samples were comparable with respect to age and sex by selecting subjects in pairs, with one member of each pair from each village, who were matched for *age and sex.*

II. PAIRED T-TEST

In **Module 29** a comparison of sample means was performed for unpaired numerical observations on the height of delivering mothers by using the **t-test.** When dealing with paired (matched) observations, comparison of sample means is performed by using a modified t-test known as the **paired t-test.**

In the **paired t-test** differences between the paired observations are used instead of the original two sets of observations.

The **paired t-test** calculates the value of t as:

$$t = \frac{\text{mean of the differences}}{\text{standard error}}$$

The degrees of freedom is the number of paired observations (= sample size) minus 1.

To interpret result of the study the same table of t-values is used as for the t-test for unpaired observations (see **Annex 29.1**).

To illustrate how the **paired t-test** is used, it will be performed on the results of the nutritional survey referred to in **Example 2** above. The results are:

Table 30.1: Results of quality control exercise during a nutritional survey

Child No.	Weight measurement (kg)		Difference
	Observer A	Observer B	A-B (kg)
1	18.6	17.7	0.9
2	17.1	14.5	2.6
3	14.3	12.4	1.9
4	23.2	20.7	2.5
5	18.4	16.8	1.6
6	14.9	14.4	0.5
7	16.6	14.1	2.5
8	14.8	17.1	-2.3
9	21.5	21.2	0.3
10	24.6	21.9	2.7
11	17.4	16.6	0.8
12	15.7	13.6	2.1
13	16.1	14.5	1.6
14	12.9	11.2	1.7
15	12.3	16.0	-3.7
16	19.4	20.4	-1.0
17	19.3	17.5	1.8
18	24.8	22.2	2.6
19	14.3	15.1	-0.8
20	13.4	10.9	2.5

The null hypothesis in this study is that if observers A and B measured all the children in the population from which these 20 children were sampled, there would, on average, be no difference between their measurements. In other words, the mean difference between A and B would be zero.

We can regard this set of 20 differences (the A - B column) as a sample of all differences that would have been obtained if the observers had measured the whole population.

To perform the significance test, the value of t has to be calculated and compared to the t-table value to determine if there is a statistically significant difference between the two. This indicates the probability that the results of the study occurred by chance.

The significance test is done as follows:

1. Calculate the mean difference of the measurements between A and B in the sample. This is the sum of the differences divided by the number of measurements:

$$\text{Mean difference} = \frac{21.1}{20} = 1.05$$

2. Calculate the standard deviation of the differences (**Module 27**):

Standard deviation = 1.77

Calculate the standard error (**Module 27**):

$$\text{Standard error} = \sqrt{\frac{\text{Standard deviation}}{\text{Sample size}}} = \sqrt{\frac{177}{20}} = 0.04$$

3. The value of t is the mean difference divided by the standard error:

$$t = \frac{1.05}{0.40} = 2.62$$

4. Refer to the table of t-values in **Annex 29.1**.

   The degrees of freedom is the sample size (the number of pairs of observations) minus 1 which in this case is 20 - 1 = 19.

   The probability from the table is < 0.05, which allows us to conclude that there is a significant difference between the observers. A and B definitely need more training and supervision on their measuring skills as the test does not show whether one or both have been inaccurate in their measurements.

III. McNEMAR'S CHI-SQUARE TEST

The **McNemar's chi-square test** is used with NOMINAL data to compare PROPORTIONS of paired observations. It is important to note that the layout of the table is different from that used with unpaired samples.

Table 30.2 shows the results of a case-control study that was conducted to determine causes of a cholera outbreak in Bombay. For each cholera case confirmed in the hospital, a subject was sought of the same sex, the same age decade and the same neighbourhood.

Table 30.2: Source of drinking water by cholera patient/control pairs in the 5 days preceding illness (incorrect layout).

	Cases	Controls	TOTAL
Shallow well	42 (55%)	15 (20%)	57
Tap water	34 (45%)	61 (80%)	95
Total	76 (100%)	76 (100%)	152

However, the layout of **Table 30.2** is not correct, as it does not take account of the fact that cases and controls were selected as pairs.

The correct layout is presented in **Table 30.3**.

Table 30.3: Source of drinking water by cholera patient/control pairs in the 5 days preceding illness (correct layout)

Controls	Cases		TOTAL
	Shallow well	Tap water	
Shallow well	12 (16%)	3 (4%)	15
Tap water	30 (39%)	31 (41%)	61
Total	42	34	76 (100%)

Adapted from Baine & Mazotti et al. (1973).

How should we interpret **Table 30.3**?

In 12 pairs both cases and controls drew water from the shallow well and 31 pairs drew water from the tap. These 43 pairs therefore give us no information whether drawing water from shallow wells is a risk factor for getting cholera or not. However, in 30 pairs (39%) the cases draw water from shallow wells, while the controls drew water from the tap, whereas in only 3 pairs (4%) the controls draw water from the shallow wells while the cases drew water from the tap. It would seem therefore that drawing water from shallow wells was a risk factor for getting cholera.

Before we accept that conclusion, we must perform a significance test to estimate the likelihood that these results are due to chance or sampling variation only. In this case the appropriate significance test is **McNemar's chi-square test** (see **Table 28.1**):

$$\chi^2 = \frac{(|r - s| - 1)^2}{r + s} \text{ with 1 degree of freedom}$$

where r = the number pairs where a control drew water from the tap and the case from the shallow well,
s = the number pairs where a control drew water from the shallow well and the case from the tap, and
$|r - s|$ means the difference between r and s as a positive number, irrespective of whether s is larger than r.

To check for statistical significance we use ordinary chi-square tables (**Annex 29.2**).

Note:

McNemar's χ^2 test is only valid if (r + s) is larger than 10.

The test can be performed on the data in our example because r + s (30 + 3) is larger than 10.

The calculation of the chi-square value is as follows:

$$\chi^2 = \frac{(30 - 3 - 1)^2}{30 + 3} = \frac{26^2}{33} = 20.5 \text{ with 1 degree of freedom}$$

Using an α-level of 0.01, the χ^2 table value is equal to 6.63 (**Annex 29.2**). We can see that the calculated χ^2 value of 20.5 is larger than the table value. This means that the p-value is less than 0.01. We therefore reject the null hypothesis and conclude that drawing water from the shallow wells was a risk factor for getting cholera.

GROUP WORK

If your data were collected by paired or matched observations, identify the appropriate statistical test and make the necessary calculations and analysis.

REFERENCES

See epidemiological and statistics textbooks referred to in **Modules 9** and **28**.

Baine WB, Mazotti M, Greco D et al. (1974) Epidemiology of cholera in Italy in 1973. *Lancet* ii (Dec.): 1370 – 1374.

Trainer's Notes

Module 30: DETERMINING DIFFERENCES BETWEEN GROUPS, Part II: ANALYSIS OF PAIRED OBSERVATIONS

Timing and teaching methods

½ hour	Introduction and discussion
2 hours	Group work

Introduction and discussion

- If none of the teams has paired observations and if participants have little experience with statistics, this module need not be presented.

- When explaining how to calculate the t-values and χ^2 values, proceed step-by-step very slowly. Again, it is more important that participants understand **how** to do the calculations, than why they are done this way.

- Take time to ensure that everyone knows how to read **Table 30.3**. This may be the first table so far in which participants deal with the numbers representing **pairs** of observations (if they have skipped the relevant part of **Module 26**).

- Make sure that participants know how to use the t-table and χ^2 table and how to interpret the results.

This page intentionally left blank

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 31

MEASURING ASSOCIATIONS BETWEEN VARIABLES: REGRESSION AND CORRELATION*

(Optional)

* Most of this module stems from the MSc course materials of the London School of Hygiene and Tropical Medicine, with permission of Richard Hayes, Betty Kirkwood and Tom Marshall

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

*These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 31: MEASURING ASSOCIATIONS BETWEEN VARIABLES; REGRESSION AND CORRELATION

OBJECTIVES

At the end of this session you should be able to:

1. **Illustrate** the relationship between two numerical variables in a scatter diagram.
2. **Interpret** a regression line.
3. **Calculate** and interpret a correlation coefficient.
4. **Perform** a test for the significance of the correlation coefficient.

I. INTRODUCTION

When exploring associations between variables we have to distinguish between nominal, ordinal and numerical data (see **Figure 28.1** and **Table 28.1** in **Module 28**).

Module 25 dealt with associations between nominal data, concentrating on case-control studies.

For associations between ordinal data, in which case **Spearman's rank correlation coefficient** or **Kendall's tau** can be calculated and tested for significance, you may refer to a textbook on statistics (see references in **Module 28**).

In this module we will examine associations between numerical data where a linear relationship is suspected.

II. SCATTER DIAGRAM

The first step in examining the relationship between two numerical variables, measured on the same subjects, is always to draw a **SCATTER DIAGRAM.**

Example 1:

In a nutrition study in a large rural district, a sample of 20 children 5 years of age were weighed and their family incomes estimated. The results were as follows:

Table 31.1: Weights and family incomes of 20 children 5 years of age

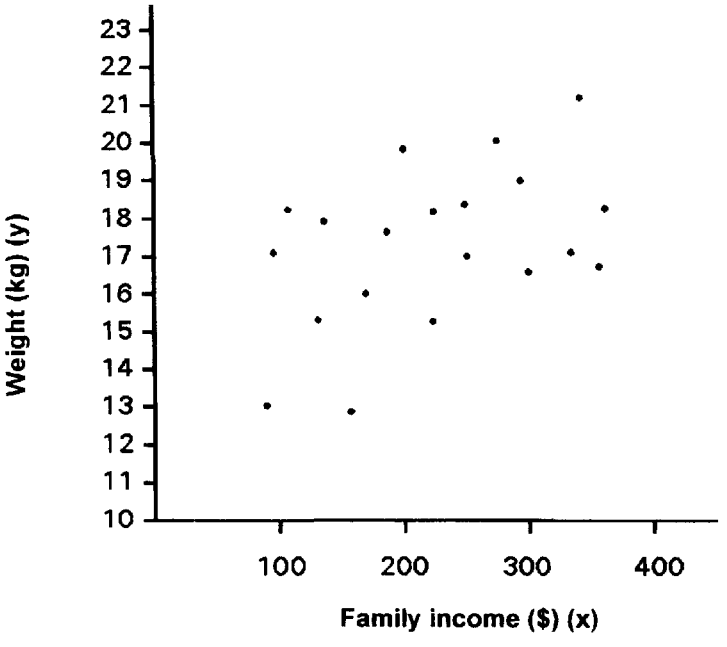
Family income in $/year	Weight in kg	Family income in $/year	Weight in kg
130	15.5	225	18.1
200	19.8	95	17.4
345	21.5	130	17.9
245	16.8	330	17.0
155	12.6	295	18.7
300	16.6	170	16.0
360	18.1	250	18.2
105	18.7	355	16.4
80	13.1	220	15.4
275	20.1	175	17.6

The objective was to examine whether, for this sample of children, weight and family income were related. It would be possible to divide the children in two income categories, a high-income category (e.g.,us$ 200 or more) and a low-income category (e.g., less than us$ 200), and to calculate and compare the mean weight in each category to see if there is a difference. One would have to use a t-test to determine if the difference is significant. (See **Module 29**.)

After performing this analysis you might conclude that children from low-income families had lower weight, on average, than children from families with high incomes. However, it would be more informative to take account of **all** the individual measurements and investigate whether the two variables 'family income' and 'weight of five-year-olds' are associated.

The following scatter diagram can be drawn:

Figure 31.1: Weights and family incomes of 20 children 5 years of age



Notes on drawing scatter diagrams:

1. If we are examining how a dependent variable is associated with an independent variable, we generally put the dependent variable on the vertical axis (the y-axis) and the independent variable on the horizontal axis (the x-axis). Sometimes it is not clear which is the dependent variable, in which case the choice of axis is arbitrary.

2. Select the scales so that the scatter fills a reasonable portion of the diagram.

3. If an axis does not start from zero, show this clearly by 'breaking' the axis (as has been done for the weight axis in the above example).

4. Label the axes clearly.

5. The plotted points should be big enough to stand out, so that the scatter is easy to look at.

III. DETERMINING LINEAR RELATIONSHIP: FITTING A REGRESSION LINE

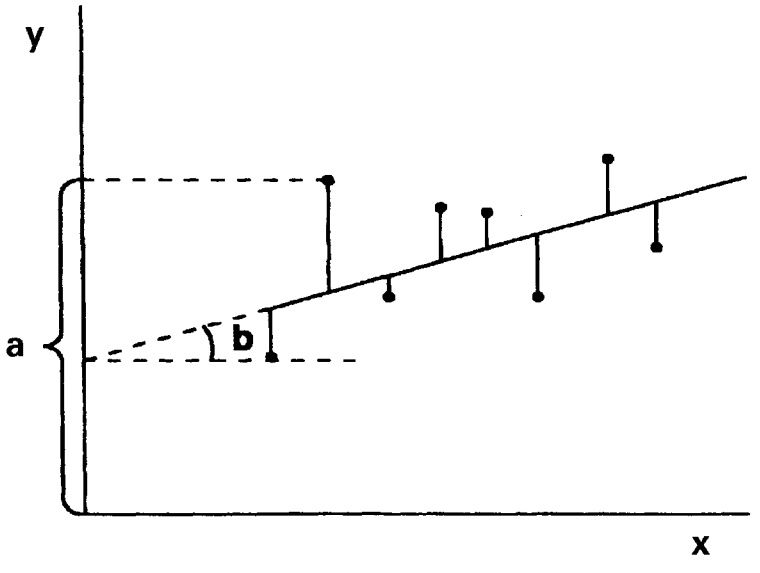
If we are interested in determining the weight of the child if we know the family income, then **weight** is the **dependent variable (Y)** and **family income** is the **independent variable (X)**.
In the scatter diagram above there appears to be an upward trend in weight with increasing family income. We can draw a line through the scatter of points, as a simple summary of the relationship between these two variables. This can be done 'by eye' - a transparent ruler is useful for this. However, fitting by eye is rather subjective and we would prefer a more objective technique. The **LEAST SQUARES METHOD** gives the 'best' straight line, using a technique that will be described below.

Any straight line drawn on a graph can be represented by the equation:

$$y = a + bx$$

Each point on the line has an x value and a y value, and the equation tells us how these x and y values are related. Different values of a and b give different straight lines. The value of a tells us the INTERCEPT of the line on the y-axis (a is the distance from zero of the point where the line crosses the y-axis, i.e., it is the value of y, when x is zero) and b indicates the SLOPE (the gradient) of the line.

To decide on the line to fit through our scatter, we have to decide what values of a and b to use. Basically we choose them in such a way that the vertical distances of the points from the line are minimised. (To be more precise, we choose a and b that minimises the sum of the squares of these vertical distances. Hence the name 'least squares method'.)



Annex 31.1 explains how the values of a and b are calculated from the data by hand. Some calculators and computer programs give the values of a and b automatically.

In our example, we find, using an appropriate calculator:

$$a = 15.09 \qquad b = 0.00984$$

So the equation of our fitted line is

$$y = 15.09 + 0.00984 \text{ times } x$$

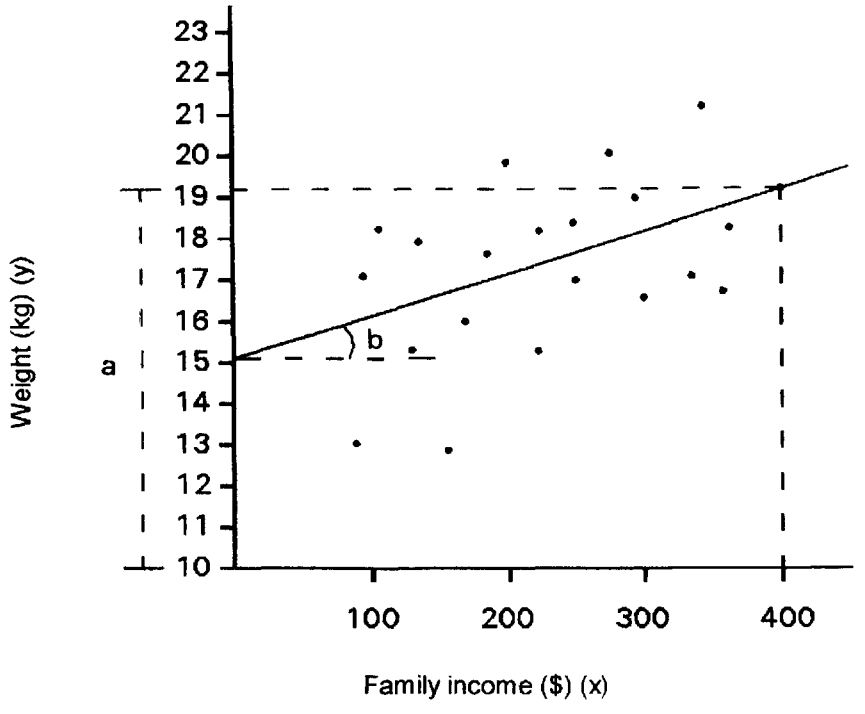
To draw this on the scatter diagram we choose two values of x, find the corresponding values of y using the equation, plot the two points on the graph and join them with a straight line.

The fitted line is called the **LINEAR REGRESSION** of weight on family income.

For example:

x = 0 gives y = 15.09
x = 400 gives y = 15.09 + (0.00984 x 400) = 19.03

Figure 31.2: Linear regression of weight of 5 years old children on family income



Interpretation of a regression line

The regression line estimates the **average** value of y for a given value of x. For example, it tells us that children whose families have an income of $200/year would **on average** weigh about 17 kg, though some would weigh more and some less than this.

The slope b, called the REGRESSION COEFFICIENT, tells us the average increase in y corresponding to a unit increase in x. So in our example, mean weight increased by 0.00984 kg (or about 10g) for each increase of $1 in family income (or about 1kg weight gain per $ 100 increase).

A note of caution:

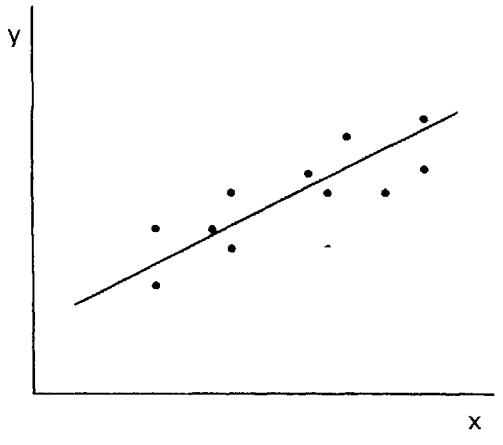
- A straight line should be fitted only if the scatter diagram suggests that the relationship between the two variables is roughly linear. More complex methods are available for fitting curves to the data.

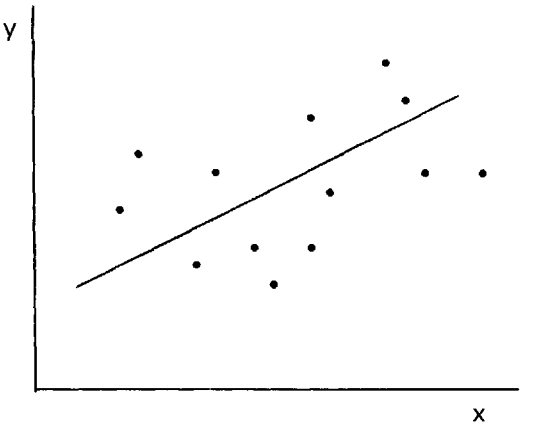
- It is dangerous to **extrapolate** the regression line outside the range of the data. In our example, extrapolating the line to an income of $ 2000/year would yield an estimated mean weight of 34.8 kg, which is of course absurd.

- In regression it is important to be clear which is the dependent and which is the independent variable, because if these were interchanged, one would get a different regression equation.

IV. CORRELATION COEFFICIENTS

Consider the following two scatter diagrams:





The regression coefficients b (i.e., the slopes of the lines) are identical in these two examples, but the scatter around the line is much greater in the second. Clearly the relationship between variables y and x is much closer in the first diagram.

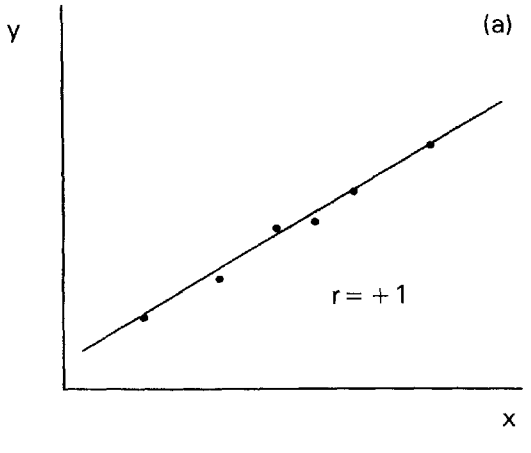
If we are interested only in **measuring** the association between the two variables, then **Pearson's Correlation Coefficient (r)** gives us an estimate of the strength of the linear association between two numerical variables.

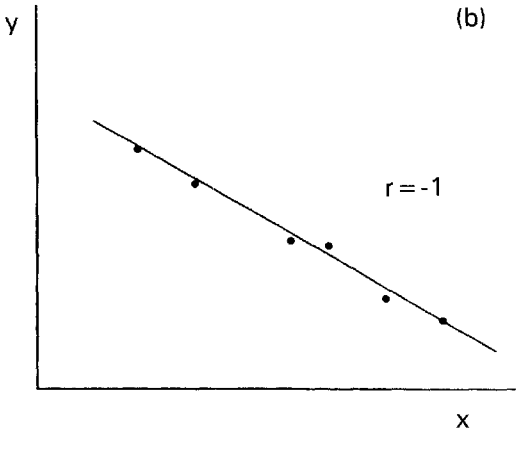
Pearson's Correlation Coefficient can either be calculated by hand (see **Annex 31.2**) or the value r can be obtained using either a calculator with built in capability to do the calculation or a variety of computer software programs.

The correlation coefficient has the following properties:

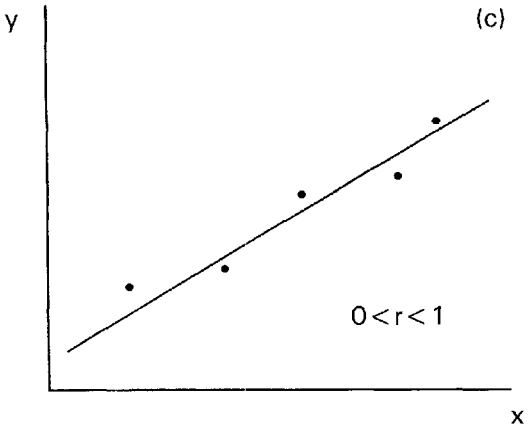
1. For any data set, r lies between -1 and +1.

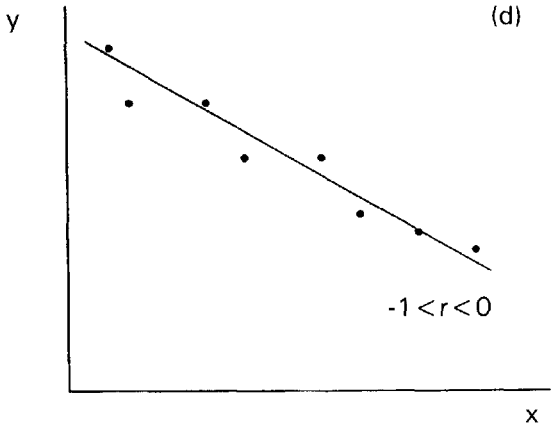
2. If r = +1, or -1, the linear relationship is perfect, that is, all the points lie exactly on a straight line. If r = +1, variable y increases as x increases (i.e., the line slopes upwards). (See **Diagram a.**) If r = -1, variable y decreases as x increases (i.e., the line slopes downward). (See **Diagram b.**)



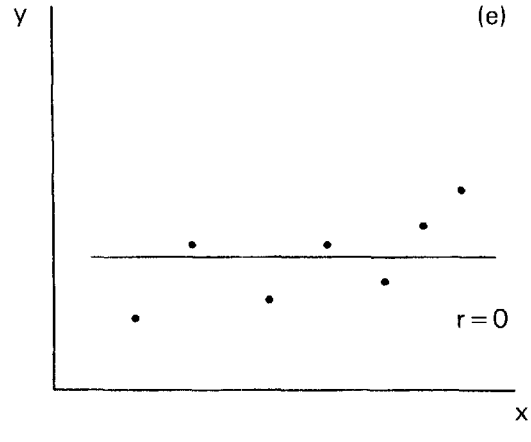


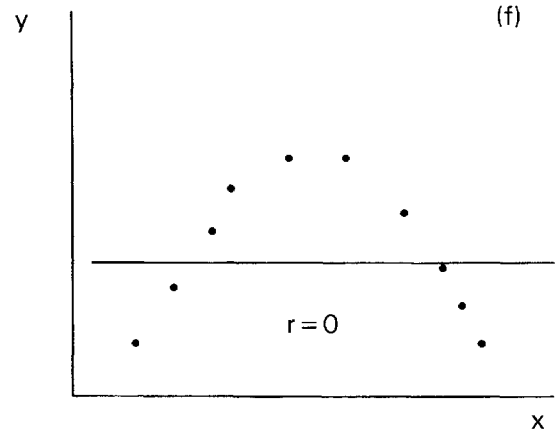
3. If r lies between 0 and +1, the regression line slopes upwards, but the points are scattered about the line. The closer r is to 1, the closer the points are to the line. (See **Diagram c.**) The same is true of negative values of r, between 0 and -1, but in this case the regression line slopes downward. (See **Diagram d.**)





4. If r = 0, there is no **linear** relationship between y and x. This may mean that there is no relationship at all between the two variables (i.e., knowing x tells us nothing about the value of y). (See **Diagram e.**) However, we could also obtain r = 0 if there were a **curved** relationship between y and x. (See **Diagram f.**)





5. A useful interpretation of r is that its square (r^2) measures the proportion of the **variability** in variable b, accounted for by the linear relationship with variable x.

Returning to our example of weight and family income, the calculator gives:

$$r = 0.414$$

which is positive (indicating an upward sloping line), but a long way from 1 (indicating that there is plenty of scatter around the line).

V. TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

The value of r was calculated from a sample of just 20 children. The result is therefore subject to sampling error and is unlikely to be equal to the true value of r, which we would obtain if we measured all 5-year-old children in this district.

The question arises as to whether there really is any relationship at all between weight and income. Perhaps in the entire population of 5-year-old children the scatter diagram would look like diagram e above (no relationship between y and x) and the positive relationship in our sample occurred by chance.

To assess whether this is the case we do a significance test on r. The null hypothesis is that in the whole population there is no linear relationship between y and x. To do the test we calculate

$$t = r \times \sqrt{\frac{n-2}{1-r^2}}$$

We compare this value of t to tables of the t distribution with (n - 2) degrees of freedom, where n is the number of observations.

In our example: n = 20 r = 0.414

Therefore $t = 0.414 \times \sqrt{\frac{18}{1 - 0414^2}} = 1.93$

Using an α-value (chosen p value) of 0.05, the t-table value for 18 degrees of freedom ($t_{18;0.05}$) = 2.10. (See **Annex 29.1**.) Thus the calculated t-value is less than the table value; this means that the p-value is larger than 0.05, and therefore the linear relationship is **not** statistically significant. Since the calculated value is close to the table value (1.93; 2.10), this is actually a 'borderline' case.

Association and causation

Note that the existence of statistical association, even if it is very strong, does NOT establish that an increase in x **causes** an increase in y, or that an increase in y causes an increase in x. A fundamental weakness of observational studies is that they can demonstrate association but not causation. To demonstrate a causal relationship one would need to choose an experimental study design.

Annex 31.1: Fitting a regression line through a scatter diagram by hand (family income and weights of children 5 years of age)

The values of a (15.9) and b (0.00984) obtained directly from the calculator can also be calculated by hand in the following way:

For the regression equation

$y = a + bx$

$$b = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n}$$

$a = \bar{y} - b\bar{x}$

Where: n is the number of observations, x is the mean of all x-values, y is the mean of all y-values.

In our example, we find, using an appropriate calculator:

Family income in $/Year ($X_i$)	**Weight in kg (Y_i)**	**X_iY_i**	**X_i^2**	**Y_i^2**
130	15.5	2015.00	16900	240.25
200	19.8	3960.00	40000	392.04
345	21.5	7414.50	119025	462.25
245	16.8	4116.00	60025	282.24
155	12.6	1953.00	24025	158.76
300	16.6	4980.00	90000	275.56
360	18.1	6516.00	129600	327.61
105	18.7	1963.50	11025	349.69
80	13.1	1048.00	6400	171.61
275	20.1	5527.50	75625	404.01
225	18.1	4072.50	50625	327.61
95	17.4	1653.00	9025	302.76
130	17.9	2327.00	16900	320.41
330	17.0	5610.00	108900	289.00
295	18.7	5516.50	87025	349.69
170	16.0	2720.00	28900	256.00
250	18.2	4550.00	62500	331.24
355	16.4	5822.00	126025	268.96
220	15.4	3388.00	48400	237.16
175	17.6	3080.00	30625	309.76
Total 4440	**345.5**	**78235.50**	**1141550**	**6056.61**

$\bar{x} = 222$, $\bar{y} = 17.275$ $\sum xy = 78235.5$

$\sum x = 4440$, $\sum y = 345.5$

$\sum x^2 = 1141550$, $\sum y^2 = 6056.6$

$(\sum x)^2 = 19713600$. $(\sum y)^2 = 119370.25$

$$b = \frac{78235.5-(4440)(345.5)/\ 20}{1141550-(19713600)/\ 20} = 0.00984$$

$$a = 17.275 - 0.00984(222) = 15.09$$

So the equation of our fitted line is

$$y = 15.09 + 0.00984x$$

Annex 31.2: Calculation of the Pierson Correlation Coefficient

The formula for calculating of the correlation coefficient is as follows:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{(\sum(x-\bar{x})^2 \sum y-\bar{y})^2}} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{(\sum x^2 - (\sum x)^2/n)(\sum y^2 - (\sum y)^2/n)}}$$

In our example of weight and family income, this would mean:

$$r = \frac{78235.5 - (4440)(345.5)/20}{\sqrt{(1141550 \quad (19713600)/20)(6056.6 - (119370.25)/20)}}$$

r = 0.414

(For interpretation, see under IV of this module.)

Trainer's Notes

Module 31: MEASURING ASSOCIATIONS BETWEEN VARIABLES; REGRESSION AND CORRELATION

Teaching and timing and methods

1 hour	Introduction and discussion
3 hours	Group work

Introduction and discussion

- This module should only be presented if at least one of the research teams needs it to analyse its data.
- It is preferable to present the correlation section first before regression, (e.g., by showing the first two scatter diagrams under section IV after **Figure 31.1,** and then continuing with section III), as participants are more likely to be able to understand it and to use it in future projects.
- Facilitators should take care to explain **when** to use correlation and when to use regression analysis.

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 32

WRITING A RESEARCH REPORT

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analyse associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analysis	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	Present summaries and draft for implementation of recommendations	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 32: WRITING A RESEARCH REPORT

OBJECTIVES

After this session, you should be able to:

1. **List** the main components of a research report.
2. **Make** an outline of your research report.
3. **Write** drafts of your report in stages.
4. **Check** the final draft for completeness, possible overlaps, and for clarity and smoothness of style.
5. **Draft** recommendations for action based on your research findings.

I. Steps in preparing the report: preliminary considerations

II. Writing the research report

1. **Introduction**
2. **Main components of a research report**
3. **Style and layout**
4. **Common weaknesses in writing**
5. **Revising and finalising the text**

I. PREPARING A RESEARCH REPORT: Preliminary considerations

Who will read your research report? How will they read it?

HSR studies have different audiences: health managers, researchers, and concerned community members. These groups will read your report from different perspectives.

Health managers and community members will ask:

- **How will this 'new information' help improve the health of the community?** (i.e., What is the problem and how will this information help solve/reduce the problem?)

Researchers, on the other hand, will want to know:

- **Can I 'believe' these findings?** (i.e., Are the findings valid and reliable?) The research design, sampling, methods of data collection and the data analysis will have to substantiate the validity and reliability.

Therefore, HSR reports should meet the needs of **health managers, researchers** and the **target group(s).**

II. WRITING THE RESEARCH REPORT

1. Introduction

HSR reports need to

- have a logical, clear structure,
- be to the point, and
- use simple language and have a pleasant lay-out.

Like an architect who designs a house has to draw a plan, you first have to make an OUTLINE for your report. This outline will contain a head, a body and a tail. The head consists of a description of your problem, within its context (the country and research area), the objectives of the study and the methodology followed. This part should not comprise more than one quarter of the report, otherwise it becomes top-heavy. The body will form the bigger part of your report: it will contain the research findings. The tail, finally, consists of the discussion of your data, conclusions and recommendations.

Then you will have to make your report attractive and user-friendly with a creative title page, a preface with acknowledgements, a table of contents, perhaps a list of tables, figures and/or abbreviations. Of course, the references you used for your study will have to be added, and annexes (including, at minimum, your data-collection tools).
Before you start writing, it is therefore essential to group and review the data you have analysed **by objective**. Check whether all data has indeed been processed and analysed as you planned in the group work of **Module 21**.

Draw major conclusions and relate these to the literature read. Again you may be inspired to go back to your raw data and refine your analysis, or to search for additional literature to answer questions that the analysis of your data may evoke.

Compile the major conclusions and tables or quotes from qualitative data related to each specific objective. You are now ready to draft the report.

2. Main components of a research report

The research report should contain the following components:

TITLE and COVER PAGE
SUMMARY OF STUDY DESIGN, FINDINGS AND RECOMMENDATIONS
ACKNOWLEDGEMENTS
TABLE OF CONTENTS
List of tables, figures (optional)
List of abbreviations (optional)

1. INTRODUCTION (statement of the problem in its local context, including relevant literature)
2. OBJECTIVES
3. METHODOLOGY

4. RESEARCH FINDINGS

5. DISCUSSION
6. CONCLUSIONS AND RECOMMENDATIONS

REFERENCES

ANNEXES (data collection tools; tables)

The findings, discussion of findings, conclusions and recommendations will form the most substantial part of your report, which has to be written from scratch.

We, therefore, strongly advise that you **start with the findings, discussion and conclusions.** Nevertheless we will briefly elaborate on each component in the sequence in which it will finally appear in your report.

- **Cover page**

The cover page should contain the title, the names of the authors with their titles and positions, the institution that is publishing the report, (e.g., HSR Unit, Ministry of Health) and the month and year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented. (See **Module 6**.)

- **Summary**

The summary should be written only *after* the first or even the second draft of the report has been completed. It should contain:

- a very brief description of the problem (WHY this study was needed)
- the main objectives (WHAT has been studied)
- the place of study (WHERE)
- the type of study and methods used (HOW)
- major findings and conclusions, followed by
- the major (or all) recommendations.

The summary will be the first (and for busy health decision makers most likely the only) part of your study that will be read. Therefore, its writing demands thorough reflection and is time consuming. Several drafts may have to be made, each discussed by the research team as a whole.

As you will have collaborated with various groups during the drafting and implementation of your research proposal, you may consider writing **different summaries** for each of these groups. For example, you may prepare different summaries for policymakers and health managers, for health staff

of lower levels, for community members, or for the public at large (newspaper, TV). In a later stage you may write articles in scientific journals. (See **Module 33.**)

- **Acknowledgements**

  It is good practice to thank those who supported you technically or financially in the design and implementation of your study. Also your employer who has allowed you to invest time in the study and the respondents may be acknowledged. Acknowledgements are usually placed right after the title page or at the end of the report, before the references.

- **Table of contents**

  A table of contents is essential. It provides the reader a quick overview of the major sections of your report, with page references, so that (s)he can go through the report in a different order or skip certain sections.

- **List of tables, figures**

  If you have many tables or figures it is helpful to list these also, in a 'table of contents' type of format with page numbers.

- **List of abbreviations (optional)**

  If abbreviations or acronyms are used in the report, these should be stated in full in the text the first time they are mentioned. If there are many, they should be listed in alphabetical order as well. The list can be placed before the first chapter of the report.

The table of contents and lists of tables, figures, abbreviations should be prepared last, as only then can you include the page numbers of all chapters and sub-sections in the table of contents. Then you can also finalise the numbering of figures and tables and include all abbreviations.

Chapter 1: Introduction

The introduction is a relatively easy part of the report that can best be written after a first draft of the findings has been made. It should certainly contain some relevant (environmental/ administrative/ economic/ social) background data about the country, the health status of the population, and health service data which are related to the problem that has been studied. You may *slightly comprise or make additions to the corresponding section in your research proposal,* including additional literature, and use it for your report.

Then the statement of the problem should follow, again revised from your research proposal with additional comments and relevant literature collected during the implementation of the study. It should contain a paragraph on what you hope(d) to achieve with the results of the study.

Global literature can be reviewed in the introduction to the statement of the problem if you have selected a problem of global interest. Otherwise, relevant literature from individual countries may follow as a separate literature review after the statement of the problem. You can also introduce theoretical concepts or models that you have used in the analysis of your data in a separate section after the statement of the problem.

Chapter 2: Objectives

The general and specific objectives should be included as stated in the proposal. If necessary, you can adjust them slightly for style and sequence. However, you should not change their basic nature. If you have not been able to meet some of the objectives this should be stated in the methodology section and in the discussion of the findings. The objectives form the HEART of your study. They determined the methodology you chose and will determine how you structure the reporting of your findings.

Chapter 3: Methodology

The methodology you followed for the collection of your data should be described in detail The methodology section should include a description of:

- the study type;
- major study themes or variables (a more detailed list of variables on which data was collected may be annexed);
- the study population(s), sampling method(s) and the size of the sample(s);
- data-collection techniques used for the different study populations;
- how the data was collected and by whom;
- procedures used for data analysis, including statistical tests (if applicable).

If you have deviated from the original study design presented in your research proposal, you should explain to what extent you did so and why. The consequences of this deviation for meeting certain objectives of your study should be indicated. If the quality of some of the data is weak, resulting in possible biases, this should be described as well under the heading 'limitations of the study'.

Chapter 4: Research findings

The systematic presentation of your findings in relation to the research objectives is the crucial part of your report.

The description of findings should offer a good combination or triangulation of data from qualitative and quantitative components of the study. There are two different ways in which you can present your findings:

(1) An integrated presentation of all data by objective

As you listed all data by objective (see **Module 21**) this should be easy.

> **For example**, in a study of factors associated with high maternal mortality, the interviews with relatives of mothers who died during delivery revealed that 90% of the mothers had attended antenatal care. Only 45% had ever given birth in a health facility, and only 25% had done so during the delivery that became fatal. Analysis of antenatal cards with the local health staff indicated that roughly 35% of mothers who visited ANC for their last baby had given birth in a hospital or HC. The analysis of cards therefore confirmed the overall low use of delivery care in a health facility, although in general the utilisation of delivery care appeared slightly better than that of the mothers who died. Interviews with the relatives of the deceased mothers presented distance and costs as major factors for not attending the delivery facilities. FGDs with mothers, however, also revealed concerns about lack of privacy in health facilities, carelessness of health staff and the young age of midwives. One woman (40 years) stated: *In that hospital, midwives are very young, they have not yet given birth themselves. How can I be delivered by my daughter, how can I let her see me in that state?*

These integrated presentations will be a compilation of tables, graphs, narrative interpretation and illustrative quotes from in-depth interviews or FGDs.

(2) Presentation of data by research instrument

Sometimes it is easier to analyse the data by instrument and integrate the findings only in the discussion. Separate analysis is indicated for objectives that are covered by distinct study populations using specific instruments.

> **For example**, in the study on reasons for non-compliance of patients with TB treatment presented in **Module 4**, four different research tools were used. Through analysis of TB cards of a cohort of patients, the proportion of irregular- and non-attendees for treatment could be defined (objective 1).

FGDs with different groups of community members provided data on community perceptions of the disease (objective 2). Interviews with patients (regular and irregular/non attendees) shed light on different reasons for non-compliance with treatment (objectives 3 and 4). Interviews with health staff revealed weaknesses in the services that could be contributing to non-compliance of patients (objective 5). Parts of the data which required integration (e.g., community members and patients had opinions on weaknesses or strengths of the services as well, whereas staff and community members complemented patient data on economic and socio-cultural reasons for defaulting) were highlighted in the discussion.

The list of data by objective will help you to decide how to organise the presentation of data. The decision concerning where to put what can best be made after all data have been fully processed and analysed, and before the writing starts.

When all data have been analysed, a detailed OUTLINE has to be made for the presentation of the findings. This will help the decision-making on how to organise the data, and is *an absolute pre-condition for optimal division of tasks among group members in the writing process.*

At this stage you might as well prepare an outline for the whole report, taking the main components of a research report (p.5) as a point of departure.

An outline should contain:

- The headings of the main sections of the report,
- The headings of subsections,
- The points to be made in each section, and
- A list of tables, figures and/or quotes to illustrate each section.

The outline for the chapter on findings will predictably be the most elaborate.

The first section under findings is usually a description of the study population. When different study populations have been studied, you should provide a short description of each group before you present the data pertaining to these informants.

Then, depending on the study design, you may provide more information on the problem you studied (size, distribution, characteristics). Thereafter, in an analytic study, the degree to which different independent variables influence the problem will be discussed.

For example: In a study on malnutrition, the chapter 'Findings' may look like this:

Chapter 4: Findings

4.1 Description of the sample

(e.g., location, age, marital status, education, soc-ec status, of mothers; age and sex of children weight/measured by research area).

4.2 Extent and seasonal variation of malnutrition in district X

4.3 Possible causes of malnutrition

4.3.1 Limited availability of food

4.3.2 Non-optimal utilisation of available food

4.3.3 High prevalence of communicable diseases

4.3.4 Limited access to MCH and weaknesses in MCH/nutrition services

This system of numbering is flexible and can be extended according to need with further headings or subheadings. It allows you to keep an overview of the process when different group members work on different sections of the report at the same time.

If your findings are very elaborate so that you have sub-sub-subheadings with 4 or 5 digits, you might decide to split up the findings into several chapters. In addition, you may consider leaving off some of the numbering on sub-sections, if it is clear under what major heading they belong. However, keep all the numbering until the final draft, as it helps you keep your report in order when various members of the group are working on different sections.

TABLES and FIGURES in the text need numbers and clear titles. It is advisable to first use the number of the section to which the table belongs. In the last draft you may decide to number tables and figures in sequence.

Include only those tables and figures that present main findings and need more elaborate discussion in the text. Others may be put in annexes, or, if they don't reveal interesting points, be omitted.

Note that it is unnecessary to describe in detail a table that you include in the report. Only present the main conclusions.

Note:

The first draft of your findings is never final. Therefore you might concentrate primarily on content rather than on style. Nevertheless, it is advisable to structure the text from the beginning in paragraphs and to attempt to phrase each sentence clearly and precisely.

Chapter 5: Discussion

The findings can now be discussed by objective or by cluster of related variables or themes, which should lead to conclusions and possible recommendations. The discussion may include findings from other related studies that support or contradict your own.

Chapter 6: Conclusions and recommendations

The conclusions and recommendations should follow logically from the discussion of the findings. Conclusions can be short, as they have already been elaborately discussed in chapter 5. As the discussion will follow the sequence in which the findings have been presented (which in turn depends on your objectives) the conclusions should logically follow the same order.

It makes easy reading for an outsider if the recommendations are again placed in roughly the same sequence as the conclusions. However, the recommendations may at the same time be summarised according to the groups towards which they are directed, for example:

- policy-makers,
- health and health-related managers at district or lower level,
- health and health-related staff who could implement the activities,
- potential clients, and
- the community at large.

Remember that action-oriented groups are most interested in this section.

In making recommendations, use not only the findings of your study, but also supportive information from other sources. The recommendations should take into consideration the local characteristics of the health system, constraints, feasibility and usefulness of the proposed solutions. They should be discussed with all concerned before they are finalised.

If your recommendations are short (roughly one page), you might include them all in your summary and omit them as a separate section in Chapter 6 in order to avoid repetition.

References

The references in your text can be numbered in the sequence in which they appear in the report and then listed in this order in the list of references (Vancouver system). Another possibility is the Harvard system of listing in brackets the author's name(s) in the text followed by the date of the publication and page number, for example: (Shan 2000: 84). In the list of references, the publications are then arranged in alphabetical order by the principal author's last name. (See **Module 5.**)

You can choose either system as long as you use it consistently throughout the report.

Annexes or appendices

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.

Examples of information that can be presented in annexes are:

- tables referred to in the text but not included in order to keep the report short;
- lists of hospitals, districts, villages etc. that participated in the study;
- questionnaires or check lists used for data collection.

Note:

Never start writing without an outline. Make sure that all sections carry the headings and numbers consistent with the outline before they are word-processed. Have the outline visible on the wall so everyone will be aware immediately of any additions or changes, and of progress made.

Prepare the first draft of your report double-spaced with large margins so that you can easily make comments and corrections in the text.

Have several copies made of the first draft, so you will have one or more copies to work on and one copy on which to insert the final changes for revision.

3. Style and layout

(1) Style of writing

Remember that your reader:

- Is short of time
- Has many other urgent matters demanding his or her interest and attention
- Is probably not knowledgeable concerning 'research jargon'

Therefore the rules are:

- Simplify. Keep to the essentials.

- Justify. Make no statement that is not based on facts and data.

- Quantify when you have the data to do so. Avoid 'large', 'small'; instead, say '50%', 'one in three'.

- Be precise and specific in your phrasing of findings.

- Inform, not impress. Avoid exaggeration.

- Use short sentences.

- Use adverbs and adjectives sparingly.

- Be consistent in the use of tenses (past or present tense). Avoid the passive voice, if possible, as it creates vagueness (e.g., 'patients were interviewed' leaves uncertainty as to who interviewed them) and repeated use makes dull reading.

- Aim to be logical and systematic in your presentation.

(2) Layout of the report

A good physical layout is important, as it will help your report:

- make a good initial impression,
- encourage the readers, and
- give them an idea of how the material has been organised so the reader can make a quick determination of what he will read first.

Particular attention should be paid to make sure there is:

- An attractive layout for the title page and a clear table of contents.

- Consistency in margins and spacing.

- Consistency in headings and subheadings, e.g.: **font size 16 or 18 bold**, for headings of chapters; **size 14 bold** for headings of major sections; **size 12 bold**, for headings of sub-sections, etc.

- Good quality printing and photocopying. Correct drafts carefully with spell check as well as critical reading for clarity by other team-members, your facilitator and, if possible, outsiders.

- Numbering of figures and tables, provision of clear titles for tables, and clear headings for columns and rows, etc.

- Accuracy and consistency in quotations and references.

4. Common weaknesses in writing

Writing is always a challenging job, which requires courage. Starting is usually most difficult. Don't be afraid to make mistakes, otherwise you will never begin! However, it is good to be aware of common pitfalls, which you might try to avoid.

An almost universal weakness of beginning report writers is **omitting the obvious.** Hardly ever does the description of the country or area contain sufficient data to permit outsiders to follow the presentation of findings and discussion without problems. On the other hand, some data (e.g., exact geographical location on the globe) could be left out which are usually in.

Endless description without interpretation is another pitfall. Tables need conclusions, not detailed presentation of all numbers or percentages in the cells which readers can see for themselves. The chapter discussion, in particular, needs comparison of data, highlighting of unexpected results, your own or others' opinions on problems discovered, weighing of pro's and con's of possible solutions. Yet, too often the discussion is merely a dry summary of findings.

Neglect of qualitative data is also quite common. Still, quotes of informants as illustration of your findings and conclusions make your report lively. They also have scientific value in allowing the reader to draw his/her own conclusions from the data you present. (Assuming you are not biased in your presentation!)

Sometimes qualitative data (e.g., open opinion questions) are just coded and counted like quantitative data, **without interpretation,** whereas they may be providing interesting illustrations of reasons for the behaviour of informants or of their attitudes. This is serious maltreatment of data that needs correction.

5. Revising and finalising the text

When a first draft of the findings, discussion and conclusions has been completed, all working group members and facilitators should read it critically and make comments.

The following questions should be kept in mind when reading the draft:

- Have all important findings been included?
- Do the conclusions follow logically from the findings? If some of the findings contradict each other, has this been discussed and explained, if possible? Have weaknesses in the methodology, if any, been revealed?
- Are there any overlaps in the draft that have to be removed?
- Is it possible to condense the content? In general a text gains by shortening. Some parts less relevant for action may be included in annexes. Check if descriptive paragraphs may be shortened and *introduced or finished by a concluding sentence.*
- Do data in the text agree with data in the tables? Are all tables consistent (with the same number of informants per variable), are they numbered in sequence, and do they have clear titles and headings?
- Is the sequence of paragraphs and subsections logical and coherent? Is there a smooth connection between successive paragraphs and sections? Is the phrasing of findings and conclusions precise and clear?

The original authors of each section may prepare a second draft, taking into consideration all comments that have been made. However, you might consider the appointment of two editors amongst yourselves, to draft the complete version.

In the meantime, other group members may (re)write the introductory sections (INTRODUCTION, OBJECTIVES and METHODOLOGY, adjusted from your original proposal).

Now a first draft of the SUMMARY can be written (see page 5 of this module).

Finalising the research report

It is advisable to have one of the other groups and facilitators read the second draft and judge it on the points mentioned in the previous section. Then a final version of the report should be prepared. This time you should give extra care to the presentation and layout: structure, style and consistency of spelling (use spell check!).

Use verb tenses consistently. Descriptions of the field situation may be stated in the past tense (e.g., 'Five households owned less than one acre of land.') Conclusions drawn from the data are usually in the present tense (e.g., 'Food taboos hardly have any impact on the nutritional status of young children.)

Note:

For a final check on readability you might skim through the pages and read the first sentences of each paragraph. If this gives you a clear impression of the organisation and results of your study, you may conclude that you did the best you could.

GROUP WORK

1. **Make an outline for your report** on a flipchart, after reviewing your objectives, your sources of information and the outcomes of your data analysis. Number proposed sections and subsections. Stick the outline to the wall in a visible place. Leave sufficient space between the lines for additions (more subsections, for example) and for changes.

2. **Start writing, beginning with the chapter on findings.** Decide with your facilitator whether you will interpret the data presenting it by variable, by objective or by study population. If you are unsure in the beginning which method of organising the presentation will work best, record your findings and interpretations by study population. In the second draft you can decide how to reorganise and shorten the presentation. **Divide writing tasks** among sub-groups of one or two persons.

3. Discuss your findings in relation to each other, to the objectives and to other literature, and write the chapter **Discussion**. Then list the major **conclusions** in relation to possible **recommendations**.

4. **Develop** at the same time **the introductory chapters** (background and statement of the problem, including new literature, objectives and methodology), adapting what you prepared for the proposal.

5. **Finally, develop the summary** following the outline given earlier in this module. Take at least half a day for this, working systematically.

6. **Keep track of progress** in writing and typing, making notes on the flipchart that has the outline of your report.

7. **Go over the first draft with the group as a whole** checking it for gaps, overlaps, etc. before the second draft is prepared. Have a facilitator from another group read the whole draft report before it is finalised.

REFERENCES

Gibaldi J (1995) *MLA Handbook for Writers of Research Papers*. New York: Modern Language Association of America.

Jen Tsi Yang et al. (1996) *An outline of Scientific Writing: For Researchers with English As a Foreign Language.* Singapore: World Scientific Publishing. www.amazon.com/exec/obidos (through internet, October 2000)

Lindsay D (1996) *Guide to Scientific Writing.* Australia: Addison & Wesley (paperback). www.amazon.com/exec/obidos (through internet, October 2000).

Trainer's Notes

Module 32: WRITING A RESEARCH REPORT

Timing and teaching methods

1 hour	Introduction and discussion
Several days	Group work

Introduction and discussion

- Put the outline for research reports on an overhead sheet and discuss it point by point. Stress that the findings, discussion, conclusions and the recommendations will have priority.
- Take an example from one of the groups when presenting a possible outline for the chapter on Findings, with appropriate headings and subheadings.
- Explain the system of numbering and differentiation of font sizes, making sure that you are consistent in the layout of headings and subheadings so that you can use the example to illustrate appropriate layout later on.
- Ask the participants to suggest the criteria they would use to judge their first draft, before you give guidelines.
- Use examples from the research proposals prepared by various groups when discussing how the statement of the problem, objectives and methodology should be adapted for the final report.

Pay attention to the need for changing the future tense used in the proposal into the present and past tense, if you suspect that some groups may overlook this aspect.

Group work

- Make sure that all of the groups first make an outline for their reports, using the outline presented in the module as a starting point. Ask the groups to hang up the outlines so everyone in their teams can see them.
- The sections on findings, discussion, conclusions and recommendations will take the most time. Some groups may find that the presentation of these sections would work best ordered somewhat differently. Let them know that they can use the outline for presentation that is most appropriate for their own data, but let facilitators discuss with their groups how this part of their presentation can be structured most logically.
- Writing can best start with the findings and conclusions. Only when a reasonable draft is ready should the research-team members be advised to (re)write the introductory chapters.
- Make sure that all group members have some writing tasks, for example, by letting them write in pairs. If certain participants have never written they might need intensive support. You can let them write several paragraphs and then rewrite the text with them, to provide them with an example.

- If groups have no experience in writing reports they will need explicit guidance concerning what points they should check on when they review their first draft, including the basic layout of the report.

- It is advisable to discuss ideas for possible recommendations during the write up of the findings and discussion. These ideas should be recorded immediately (preferably on a flip chart) so they can be used when phrasing recommendations.

- All facilitators should comment on the summary section prepared by each group. Each facilitator should also read and comment on at least one complete draft report from another group before it is finalised.

Designing and Conducting Health Systems Research Projects
Part II: Data Analysis and Report Writing

Module 33

DISSEMINATION, COMMUNICATION AND UTILISATION OF RESEARCH FINDINGS

Steps in data analysis and report writing

Questions you must ask	Steps you will take*	Important elements of each step
What data have been collected for each research objective? Are data complete, accurate?	Prepare data for analysis	Review field experience Make an inventory of data for each objective/study population Sort data and check quality Check computer outputs (21)
What do the data look like? How can the data be summarised for easy analysis?	Summarise data and describe variables/identify new variables	Frequency tables, figures, means, proportions, descriptive cross-tabulations, (*quantitative data*) (22, 24); Coding, listing, summarising data in compilation sheets, matrices, flow charts, diagrams and narratives (*qualitative data*) (*23*)
How can the associations between variables be determined?	Analysis associations	Analytic cross-tables (24) Measures of association based on risk(25) Dealing with confounders (26)
	Prepare for statistical analyse	Measures of dispersion, Normal distribution and Sampling variation (27)
Do we measure differences or associations between variables?	Determine the types of statistical analysis	Choosing significance tests (28)
How can differences between groups be determined?	Analyse unpaired and paired observations	t-test, chi-square test (29) ** paired t-test, ** McNemar's chi-square test (30)
How can the associations between numeric variables be determined?	Implement measures of association	** Scatter diagram, ** Regression line and ** Correlation coefficient (31)
How should the report be written?	Write the report and formulate recommendations	Prepare outline for report Present and interpret data Draft and redraft Discuss and summarise conclusions Formulate recommendations (32)
How should the findings and recommendations be communicated, disseminated and used?	**Present summaries and draft for implementation of recommendations**	Discuss summaries and plan for implementation with all stakeholders (33)

* These steps need not be in the sequence in this diagram. **The sequence may be adjusted according to the needs of the research teams.**

** These elements are optional and may be omitted if not relevant for research teams

Module 33: DISSEMINATION, COMMUNICATION AND UTILISATION OF RESEARCH FINDINGS

OBJECTIVES

After this session you should be able to:

1. **Develop a strategy** for the dissemination, communication and utilisation of your research findings.
2. **Prepare a presentation** of your research findings for stakeholders.
3. **Prepare a plan of action** for promoting the utilisation of your research recommendations.

I. Introduction

II. Strategy for the dissemination, communication and utilisation of research findings

III. Presenting the research findings to different stakeholders

IV. Preparing a plan of action for promoting the utilisation of your research recommendations

I. INTRODUCTION

Even the greatest research findings mean very little unless they are effectively disseminated, communicated and used. The beneficiaries of research are not only the health professionals. There is a need to promote and increase the utilisation of your research results among all potential users, varying from community members to donor agencies. The likelihood of research findings being used will increase if the following steps are taken:

KEY STEPS

(1) Develop and use a systematic dissemination and communication strategy for reaching different audiences of potential users;

(2) Present the research results to all stakeholders and obtain feedback on findings and recommendations; and

(3) Develop a plan of action to promote the implementation of the recommendations that resulted from your study.

II. STRATEGY FOR DISSEMINATION AND COMMUNICATION OF RESEARC RESULTS

The purpose of HSR is to provide useful information to managers *at all levels* that will facilitate problem solving. Therefore an extremely important step in the health systems research process is the presentation of the research results to all interested parties so you can discuss with them the findings, recommendations and possibilities for action. Merely producing and disseminating a report or a research paper is usually not adequate.

A strategy for dissemination and communication to promote utilisation of research findings should be developed, taking into account the following **elements:**

1. Recapitulate the problem, the major contributing factors and proposed actions to solve it.

This summary will focus your attention on what you hope to achieve with your study and help you determine to whom you will present the results of the study, in order to mobilise them for action.

> **For example**, in a study of factors contributing to frequent cholera epidemics in Kabwe District, Zambia, a major contributing factor is poor human waste disposal shown by 10% latrine coverage. The results need to be presented to village leaders and health committees, at minimum. A major recommendation is to increase the latrine coverage. Possible action is a participatory latrine construction programme.

2. Identify different partners (groups or institutions) and their potential contributions for solving the problem.

You have to clearly identify who forms the target group for action, which institution will be responsible for implementation, which staff will carry out the day-to-day activities, which politicians should give their blessing, and who can provide financial support. All those groups and institutions will have to be informed about the results of your study; they should be able to react to your findings and recommendations and commit themselves to some form of action.

It is always advisable to start finding support 'close to home'. The Kabwe research team that carried out the study partly consisted of District Health Management Team (DHMT) members. They distinguished the following institutions and persons for feedback:

- **Responsible institution:** District health management team (DHMT)
- **Target group:** Village leaders and village health committees. Ultimately: heads of households
- **Technical support:** MOH Community Water and Sanitation Programme; NGO supporting water and sanitation activities
- **Political support:** District medical officer (DMO, chair person of the DHMT)
- **Financial support:** DHMT + NGO/donor concentrating on water and sanitation
- **Day-to-day implementation:** Health assistants supporting heads of households

In this case, the DHMT is proposed as the crucial partner from the government side. The DMO should provide political support for further fundraising; the district health officer (DHO) who was part of the research team is responsible for technical implementation. He has links with the manager of the MOH Community Water and Sanitation Programme, which needs to provide technical support, and with the health assistants in the field who fall under his authority.

The heads of households are the major partners in the community. They can be best approached through the village leaders and village health committee (or village development committee, if there is no strong VHC). However, village health workers, if existing, will also be important partners, like women's groups, as women are a major group of users and maintainers who are likely underrepresented among the heads of households. The leaders of the church or mosque can be co-operative mobilisers as well. The support of a local district council member or MP might likewise be useful.

Identification of financial partners is crucial for implementation of research recommendations. Resources should first be found in the regular budget of the district authorities and Ministry of Health. Usually some additional funds are required. In Kabwe district, an NGO could be identified which was specialised in the field of water and sanitation and willing to provide technical as well as financial support.

At this stage it is also wise to consider **possible barriers** to the proposed action. Sometimes there are groups in the community who are less interested in an activity or even oppose it. These groups should get special attention in the feedback process of the research results.

> **For example**, in a study on factors contributing to late reporting for treatment of leprosy patients in Aceh, Indonesia, it was found that many patients followed traditional treatment before they came with their symptoms to a Health Centre. Neither traditional healers (*dukun*) nor community members were sufficiently aware of the availability of powerful short course treatment. When the results of the study were fed back to the community leaders and possibilities for intensified case-finding were discussed, care was therefore taken to invite the local *dukun* as well.

3. Select appropriate information channels

Partners identified should normally be informed through **interpersonal communication**. In case of the Kabwe study, the DHMT would be best approached at a *regular meeting*. Similarly, a meeting of village leaders could be easily convened, together with VDC and VHC members, VHW religions leaders and representatives of women's groups. *Personal briefings* would have to be organised between research team members and the MOH Water and Sanitation Programme manager. The NGO would have to be approached separately as well.

Now, or later, when action is planned, one should also think of the **mass media** as information channel: *newspapers, radio, TV, posters.* These are effective in disseminating information to a wide range of audiences. Mass media and interpersonal channels may fruitfully complement each other. Identify key contact persons in each news organisation, for example reporters who cover the area of your research, and/or the news editor. You may approach them through a letter including relevant information about your project. Existing summaries of the study may be used, but it might be useful to elaborate on the problem so that the reason for the study can be fully understood by a non-health person.

In addition, other researchers may be interested as well in the research results and research methods used, especially those in community health or social science departments. Articles in *research journals* and presentation of research results to interested students and staff are therefore appropriate, even if those presentations are not directly related to action. It may also be useful to present results from HSR projects at *'research days'* organised to improve communication between a selected group of health managers and health researchers.

Annex 33.1 provides an overview of a strategy for dissemination and communication of research findings based on the example of the human waste disposal study in Kabwe district.

III. PRESENTING THE RESEARCH RESULTS TO DIFFERENT STAKEHOLDERS

Having identified the most likely partners for implementation of the research results and the channels through which you will approach them, it is now time to consider the presentation itself. A number of points should be kept in mind.

1. **Make sure that sufficient time is allocated for the presentation and for discussion.**

   - Prepare a presentation;
   - Ensure sufficient time for presentation (some 15-20 minutes at least!);
   - Allow as much time for discussion as possible to get feedback on your findings and recommendations; and
   - Include discussion on the next steps (action).

2. **Arrange your presentation.**

Your presentation may consist of:

- A brief introduction, including the statement of the problem, objectives of the study, sample(s) and data collection tools used.

- The major findings, listed in a logical sequence (for example, starting with a description of the problem, followed by the major variables which influence it).

- The recommendations, roughly following the same sequence.

Preferably separate recommendations for policymakers and health managers, for health staff and for community members. This makes it easy to talk with each group.

For example, feedback on the Kabwe study to *community members* would have to stress the likelihood of getting cholera through contact with human waste, and the role of latrines in cholera prevention. This information should come as background to data about the magnitude of the cholera epidemic. The low latrine coverage and the reasons for this low coverage resulting from the study, would need to be thoroughly discussed in order to get suggestions for a participatory latrine construction intervention. Feedback to the *DHMT*, the *Water and Sanitation Programme Manager* and the *NGO* should stress the association between cholera prevalence and poor waste disposal in Kabwe District, the reasons for the low latrine coverage and the reaction of community members on the proposition to start a participatory latrine construction intervention.

Remember that your audience basically wants to know: 'How can we solve this problem?' Therefore:

- Avoid technical jargon.

- Do *not* overload the audience with statistical data. However, you might present some tables to support your main conclusions, and illustrate the problems you identified with some interesting observations.

- Be *specific* in your recommendations concerning the actions required to solve the problem. If your research indicates that there are several viable options, describe the alternatives and their potential advantages and disadvantages, and give ample opportunity for discussion. Be careful not to make the impression that you are 'telling your audience what to do.'

3. Prepare appropriate visual aids

Have sufficient copies of your 'Summary of findings and recommendations' for all who are present. If the presentation takes place after you have completed your report, have some copies of the full report available for those who are most concerned or interested.

Prepare overhead sheets, slides or flipcharts to highlight the most important points in your presentation (e.g., problem, main objectives, major findings and recommendations that require action from those you are addressing).

4. Discuss the findings and the logic and feasibility of the recommendations with different target groups

Check whether they agree with your conclusions concerning the nature, magnitude and causes of the problem, based on their own experiences. Solicit additional information on questions that remain. Concentrate on the discussion of recommendations that concern them, to obtain their opinion on appropriateness and feasibility and elicit their support for any actions they themselves should take. Briefly inform them about the recommendations to other stakeholders and their reaction, and solicit additional suggestions for action.

5. Appoint two team members as recorders for each session.

Make sure that proper minutes of the discussion are taken, especially concerning the decisions and follow-up actions that are agreed upon. These minutes should subsequently be circulated to all those who were present, as well as to key persons who were invited but did not personally attend.

GROUP WORK, PART I

1. Prepare presentations for selected groups. Allow sufficient time for developing the presentation - at least half a day for drafting the outline of what you will present and any notes you will need and then adapting your 'Summary of major findings and recommendations' for distribution, if necessary. Reserve another half a day for inserting corrections, making copies of the summary and any handouts and preparing overhead sheets or flipcharts.

2. The team leader or the most articulate member of the research team may make the entire presentation, but it is also possible to assign various parts of the presentation to different group members.

IV. PREPARING A PLAN OF ACTION

When the presentations have led to positive responses, the moment has come to **draft a plan of action** for the coming year. In this plan, the strategy framework discussed in the first section of this module needs to be elaborated, incorporating the feedback received from the community, relevant health managers and the potential donor/NGO.

If there are a number of specific actions to be taken for which detailed planning by several parties is needed, you might consider **holding an 'action planning workshop'**. In this setting the various groups involved could work together to study the findings and recommendations in detail and develop an action plan. The presentation for managers and community leaders could be expanded to include this active planning phase, or a separate workshop could be scheduled, after key decision makers have had time to review the results of the study and their implications.

The workshop may vary from a couple hours to a couple days in length, depending on the size of the study and the nature of the actions that need to be planned. If a longer format is chosen, you might consider a short field visit to some of the research sites before the group begins the 'action planning phase' of the workshop.

When plans are drafted, be sure the working groups consider what activities and tasks will be completed, who will be responsible, when they will take place, and what resources are needed.

We will take the recommendation to start a **participatory latrine construction intervention** from the Kabwe study as an example of how such an action plan could be developed.

Assuming that all approached parties (community members, DHMT, Community Water and Sanitation Programme, NGO) have agreed to participate, the details of this participation would have to be specified. It could be proposed that the research team, backed up by the DHMT, organise the following one-day workshops:

(1) One with representatives of the Community Water and Sanitation Programme, of the NGO, the district health officer and some health assistants, to discuss technical details and costs of different types of latrines.

(2) One with different community leaders, (+ representatives of VDCs, VHWs, women's clubs, and religious leaders) about wishes and possibilities with respect to latrine construction in the community, followed by similar meetings in all wards of the district.

(3) One with all partners about final target, type(s) of latrines and rough estimation of costs, material contributions of all partners and further supportive action required.

The meetings should cover the following action points:

1. First of all, the **target of the intervention** would have to be specified. For example: an increase in latrine coverage in Kabwe District from 10% to 80% of all households within three years.

2. Then the material **contribution required from all partners** needs specification. Heads of households should be prepared to provide labour and part of the building materials and equipment; the district health management team should be willing to incorporate the latrine construction programme in its environmental health activities, providing manpower (health assistants, district health officer) and transport; the interested NGO could be asked to support the heads of households with additional building materials and to support the DHMT by paying for the additional transport and costs for training of health assistants and village health workers. Moreover, the NGO could be asked to provide yearly incentives for the 25 best-constructed and 25 best-maintained latrines in the district.

3. **Further institutional supportive action** needs to be specified at all levels. At **community level**, the VHC should be willing to take responsibility for mobilising the households for latrine construction and maintenance (e.g. in village meetings, through women's clubs etc.). Its members could be assisted by the youth club of the church, through drama performances on hygiene and the role of latrines in the control of cholera and other diseases. The VHC could also support VHWs who would be trained to provide technical advice on latrine construction and supervise and monitor the latrine construction in the village. At **DHMT level** the District Environmental Officer would have to make a detailed work plan and budget for three years, including all mobilisation, training, construction and supervision, monitoring and evaluation activities. The **MOH Water and Sanitation Department** and the supporting **NGO** should be willing to assist the DEO in the development of simple *building instruction sheets* for household level, *training materials* for Environmental Health Assistants and *IEC materials* for schools (provided schools are willing to participate).

   The **mass media** could be approached to provide more information about the intended participatory latrine construction intervention as soon as the first village has started with the construction of latrines. One could think of a *newspaper article* with photographs, and a *radio or TV broadcast* about the problem (cholera, low latrine coverage), the planned intervention and the first results.

Annex 33.2 provides a schematic summary of the proposed action plan

GROUP WORK, PART II

1. Prepare a plan of action for implementing your study recommendations;
2. If the plan for implementing results should involve the participation of several individuals or organisations, consider whether an 'action planning workshop' would be useful.
3. If so, make a specific plan for it. Send the notes of the previous discussions around, make objectives for the meetings and fix a date in consultation with the partners invited.
4. During the workshops, try to be as detailed as possible about the nature and timing of the contributions and make a feasible work plan for the different activities. The DHMT, in particular the person responsible for implementation of the action plan, after the workshops should prepare a more refined work plan for a specific period (1-3 years) combining the tentative work plans for different activities.

Trainer's Notes

Module 33: DISSEMINATION, COMMUNICATION, AND UTILISATION OF RESEARCH FINDINGS

Timing and teaching methods

30 minutes Introduction and discussion

Introduction and discussion

It is important to present this module before participants start preparing the summary of their findings and recommendations.

Note:

Stress that the ACTION PLAN should consider how to mobilise critical stakeholders for the problem researched.

Group work

When presenting the research results it is recommended that **participants distribute the summary including the full recommendations to all participants and guests of honour**, to enhance their participation in the discussion. Details on the specific objectives of the study, the methodology, and the findings (for example, some crucial tables or graphs) can be presented using flip charts or overhead sheets.

Annex 33.1: Strategy for dissemination and communication of research findings to promote utilisation (study on human waste disposal, Kabwe district, Zambia)

Major problem	Major proposed action	Potential partners	Information channels	Instruments and major discussion points	Expected outcome
Cholera epidemics	Participatory latrine building programme	**Community:** - Village leaders - Village Health Committee - VH Workers - Women groups - Religious leaders	Village meetings	Report summaries Overheads Posters Oral explication Discussion	Agreement for - Participation in construction, maintenance and use of latrines
Poor human waste disposal (10%latrine coverage)		**MOH** - District Health Management Team - Community Water and Sanitation Programme Manager - Environmental Health Assistants	Regular meetings Personal briefings Meeting with District Environmental Officer	- Poor hygiene conditions cause spread of cholera - Necessity for increase of latrine coverage	- Political and technical support; - Necessary resources (manpower, some money and equipment)
Lack of knowledge lack of support		**NGO** interested in water and sanitation	Personal briefings	- Possibilities for participation in latrine construction programme	- Additional technical and financial support
		Mass media	Personal briefings		- Newspaper article on research findings and planned latrine building programme

Annex 33.2: Summary action plan for participatory latrine construction intervention, Kabwe district, Zambia)

Agreed target	Partners	Material contribution	Further institutional supportive action	Expected outcome	Impact
Increase in latrines from 10-80% over 3 years	**Community** - Village leaders - VHC and VHW - Women groups - Church leaders	**Heads of households** - Labour - Part of (local) materials and equipment	**Community** - Mobilisation and supervision by VHC - Drama by church youth club - Technical advice by VHW	- Increased construction of latrines from 10-80% - Increased maintenance and use of latrines - Better hygiene and sanitation behaviour	Strong reduction in severity and frequencyof cholera (ultimately: elimination!)
	MOH - DHTM - Community Water and Sanitation Programme - Environmental Health Assistants	**MOH/DHMT** - Manpower (Env. Health Assistants + DHO) - Transport (partly) - Environmental Health Assistants	**DHMT/DHO** (supported by CWSP) - Detailed workplan for 3 years with detailed budget - Simple building instructions latrines - Training of Env. Health Assistants + VHW - Regular supervision, monitoring and evaluation		
	NGO interested in water & sanitation	**NGO** - Additional transport + training costs - Additional building materials - Yearly incentives for 25 best constructed and 25 best maintained latrines	**NGO** - Assistance in development simple building instructions, training materials for VHW & EHAs, IEC		
	Mass media		**Mass media** - Materials for schools, - Newspaper articles, and - Radio broadcasts, on planned and implemented activities		

ABOUT THE AUTHORS

Corlien M. Varkevisser, MA, PhD, MPH, is a medical sociologist-anthropologist by profession who specialised in public health. As a staff member of the Royal Tropical Institute, Amsterdam, and former head of the Primary Health Care (PHC) Unit, she has gained extensive experience in health systems research and PHC management in sub-Saharan Africa. She was one of the co-initiators of the Joint HSR Project (WHO/Netherlands Ministry for Development Cooperation/ Royal Tropical Institute) for Southern Africa and was based at the WHO Sub-regional Office in Harare as manager of the Joint HSR Project from its onset in April 1987 till 1992. Thereafter she became manager of the MPH course at the RTI and professor in HSR at the Faculty of Political and Social Sciences, University of Amsterdam.

Indra Pathmanathan, MMBS, MPH, is a physician specialised in public health who, as Head of the HSR program of the Ministry of Health in Malaysia since its inception, has been responsible for developing and implementing several strategies for HSR that have been replicated in other countries. These included training programs in HSR and Quality Assurance for decision-makers in ministries, for physicians, and for staff in district health teams, hospitals, and universities. She was a member of the Advisory Group on HSR, WHO-Geneva and served on the editorial board of BRIDGE. Over the past ten years she has been a consultant to the World Bank in the field of health.

Ann Brownlee, MA, PhD, is a medical sociologist who specialized in HSR, planning and evaluation, and cross-cultural aspects of health care. She served as Research and Evaluation Coordinator for the Project for Strengthening Health Delivery Systems in West and Central Africa for a number of years, where she worked closely with WHO's Regional Office for Africa and with colleagues from Africa and elsewhere to develop an HSR training program and to publish the *HSR Training Course* that was a forerunner of this volume. She currently works as a consultant in international health for groups such as WHO, IDRC, and Wellstart and teaches at the University of California at San Diego.