These study materials for the distance learning Epidemiology course have been prepared by the London School of Hygiene & Tropical Medicine (LSHTM).

Dean of Studies: Sharon Huttly
Course Directors: Anne Tholen and James Hargreaves (Course Content Director)
Head, Distance Education & Professional Development Office: Sue Horrill

Acknowledgements
LSHTM would like to thank all the staff and associates of the School who developed and wrote these materials.

Any comments on this study pack, favourable or unfavourable, would be most welcome and should be addressed to:

Head, Distance Education & Professional Development Office
London School of Hygiene & Tropical Medicine
Keppel Street
London WC1E 7HT
United Kingdom

Telephone: +44 (0)20 7299 4658
Email: dlsupport@lshtm.ac.uk

© University of London 2011

The University has made every effort to trace copyright holders. We apologize for any omissions and will welcome additions or amendments for inclusion in any reprint. This material is not licensed either for resale or for further copying.
## Contents

### Statistical methods in epidemiology

<table>
<thead>
<tr>
<th>Information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Module introduction</td>
<td></td>
</tr>
<tr>
<td>Module specification</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stata glossary</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of datasets</td>
<td></td>
</tr>
</tbody>
</table>

### Section 3: Practical sessions

<table>
<thead>
<tr>
<th>Practical 1</th>
<th>Measures of effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practical 2</td>
<td>Cohort studies</td>
</tr>
<tr>
<td>Practical 3</td>
<td>Survival analysis</td>
</tr>
<tr>
<td>Practical 4</td>
<td>Case-control studies</td>
</tr>
<tr>
<td>Practical 7</td>
<td>Logistic regression 1</td>
</tr>
<tr>
<td>Practical 8</td>
<td>Logistic regression 2</td>
</tr>
<tr>
<td>Practical 9</td>
<td>Logistic regression 3</td>
</tr>
<tr>
<td>Practical 10</td>
<td>Matched case-control studies</td>
</tr>
</tbody>
</table>

### Section 4: Solutions to the practical sessions

<table>
<thead>
<tr>
<th>Solutions to Practical 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Solutions to Practical 2</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 3</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 4</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 7</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 8</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 9</td>
<td></td>
</tr>
<tr>
<td>Solutions to Practical 10</td>
<td></td>
</tr>
</tbody>
</table>

### Reader

<table>
<thead>
<tr>
<th>Reader contents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader articles</td>
<td></td>
</tr>
</tbody>
</table>
Module Introduction

The London School of Hygiene & Tropical Medicine welcomes you to:

**EPM202 Statistical Methods in Epidemiology**

We hope you enjoy studying this module.

---

**Contents**

Finding out about the module ................................................................. 1
The module in context ............................................................................. 2
Module calendar ...................................................................................... 2
CAL (Computer Assisted Learning) material ............................................. 2
Workbook ............................................................................................... 3
Readings .................................................................................................... 4
Assignments ............................................................................................. 5
Web conferencing .................................................................................... 5
Website ..................................................................................................... 6

---

**Finding out about the module**

Before you start going through the module material, we recommend you get an overview of this module and how it is run. To do this, we suggest you first read the **Module Specification** document which is provided with your course materials and may also be downloaded from the course website. It provides an at-a-glance source of key information about the module such as:

- The title and course code for the module (sections 1 and 2).
- The overall aim of the module and its learning objectives (sections 10 and 11).
- The module content (section 12).
- The learning methods used (section 13).
- The study resources provided (e.g. LSHTM materials, software, textbooks) you will need to complete the module (section 14).
- How learning is assessed (section 15).

The Module Specification should be read alongside this **Module Introduction** which gives guidance on how to go about studying this module. We recommend that you spend some time acquainting yourself with both these documents before you start working through the computer-based sessions. It is also important to
check the messages posted on the EPM202 NoticeBoard conference for key information (see under Web conferencing).

The module in context

This module is designed to provide you with the key statistical knowledge and skills you will need to analyse and interpret data from the common forms of epidemiological studies.

In the first half of the module the focus is on issues specific to different types of study. The second half of the module deals with statistical modelling and multivariable analyses. The combined materials will enable you to choose and use the techniques appropriate for estimation and hypothesis testing in selected situations. You may find some of the material in this module hard-going but you are sure to find the effort is worthwhile – and by the end of the module you will possess a comprehensive knowledge of statistical analyses and modelling techniques at a level desired by all epidemiologists! Although we have tried to avoid unnecessary mathematical detail, a proper understanding of the methodology does require the use of some mathematical formulae. No knowledge of calculus will be required.

Module calendar

<table>
<thead>
<tr>
<th>Month</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>October</td>
<td>Tutoring support begins (web conferencing, email queries, assignment marking).</td>
</tr>
<tr>
<td>January</td>
<td>AA web conference opens.</td>
</tr>
<tr>
<td>28th February</td>
<td>Final submission date for the Formative Assignment (FA).</td>
</tr>
<tr>
<td>31st March</td>
<td>Final submission date for the Assessed Assignment (AA).</td>
</tr>
<tr>
<td>April</td>
<td>Exam practice WebBoard conference open 6 weeks prior to the exams.</td>
</tr>
<tr>
<td>June</td>
<td>Exam usually held in June.</td>
</tr>
</tbody>
</table>

CAL (Computer Assisted Learning) material

- The CAL sessions form the basic learning material, and are provided on CD-ROM. If you have not yet received your CD-ROM, or do not have access to these, you may download the CAL material from the link below. Please note you will only be able to download these materials twice. http://www.lshtmldownloads.co.uk/.

- You should work through all the CAL sessions and try to ensure you understand each step before you go on to the next, as the course material builds on itself.
• Within the earlier sessions you will review some of the material you have learnt in the core modules EPM101, EPM102 and EPM103. Individual module students should be familiar with basic concepts in statistics and epidemiology before taking this module and may download the sessions from these core modules if they wish.

• We also have ‘masterfiles’ for all sessions. These are sessions in PDF format on the module CD-ROMs. The masterfiles hold the same information as the CAL material but can be printed out. Please note that masterfiles are not interactive and do not replace the CAL material. When studying the CAL sessions, we recommend that you work through the CD-ROM material first. The masterfile document can then be used for revision purposes to refer back to specific sessions if you wish.

**Workbook**

• The EPM202 Workbook accompanies the CAL sessions, and uses the statistical software package Stata. You should use the Workbook in parallel with the CAL material, so that you gain practical experience as you go along. In most of the practicals, the dataset and analysis is the same as that used for illustration in the corresponding CAL session. It may be useful as you work through the practicals to refer back to the corresponding CAL session for in-depth details about the methods.

• The Workbook is divided into four Sections:
  - Section 1 contains a Stata glossary of the commands used in statistical analysis, including those applied in the practical sessions.
  - Section 2 describes the various datasets used in this module.
  - Section 3 contains all the practical sessions. Each of these sessions consists of a tutorial which guides you through the analysis followed by a review exercise where you will ‘have a go on your own’.
  - Section 4 has the solutions to the review exercises.

• The practical sessions give you the opportunity to gain experience in performing your own analyses. The interpretation of the output is as important, if not more important than being able to use Stata to produce the output. Each practical session consists of a tutorial type exercise which guides you through the analyses. In the tutorials you will be instructed as to which commands to use and provided with interpretation of the output. At the end of each tutorial there is a review exercise where you will ‘have a go on your own’ and apply the commands used in the tutorial. The solutions to these review exercises are given in Section 4 of this Workbook.

• The example datasets used in the practical sessions all come from studies in which staff of LSHTM been closely involved. Although these datasets may not be appropriate for your work, we hope that you will find that they provide meaningful examples of specific study designs to which the methods are applied. A description of each dataset is given in Section 2 of this Workbook.
Module introduction

Before starting these practicals, you must copy the required Stata commands and datasets from the EP202 CD. On this CD there is a folder called \Stata. In this you should find folders called 'ado' and 'data'. You must copy all of the files that are in these folders into the corresponding folders on your C: drive (or whatever drive letter you use). That is, copy all files from \Stata\ado on the CD into C:\ado and copy all files from \Stata\data on the CD into C:\data. These C:\ado and C:\data folders are the default locations in which Stata looks for its 'ado' and 'dta' files. Once you have copied the files you should find that the instructions work.

Be especially careful that you do not duplicate folders, for example, by putting the files into C:\ado\ado. This is a common error and it means that the commands will not work.

It is essential that prior to working through a practical session:
- you have studied the corresponding EPM202 CAL session
- you have a basic knowledge of Stata
- you have copied your EPM202 datafiles into an appropriate workspace.

If you wish to refresh your knowledge of Stata refer back to core module EPM102 Statistics with Computing.

Please note this Workbook uses Stata 10 commands. While there are some differences between Stata 10 and 11, all commands used in the Workbook will also work in Stata 11.

Please note: not all the CAL sessions have a practical associated with them.

Readings

Three papers are included in the Reader, which illustrate the application and presentation of the statistical methods discussed within the module.

You should make use of the textbook supplied with your material (Essential Medical Statistics).

Other books recommended as optional reading for this module include (these books are not supplied):
- Multivariable Analysis: A Practical Guide for Clinicians (Katz and Mitchell)
- A Short Introduction to Stata for Biostatistics (Hills and de Stavola, pub: Timberlake).

Please also make use of the Stata Guidelines - A Stata Summary for DL Students - a summary of all the commands you will meet in the EP statistical modules. This can be downloaded from the General Resources page on the student website http://dl.lshtm.ac.uk/programme/student/ep/student/general.htm.

We recommend you also make use of the LSHTM on-line library resources (access via the University of London International Programmes portal). Details on how to access the library are given in the Student Handbook.
Assignments

- **Formative Assignment.** We recommend that you complete the Formative Assignment (FA). This tests your understanding and the feedback given by tutors (and specimen answer) enables you to see how you are progressing. The FA will also help you prepare for the Assessed Assignment (AA). The FA submission deadline is 28th February.

- **Assessed Assignment.** The Assessed Assignment (AA) is a compulsory element of the module, comprising 30% of the final module grade and you must obtain a minimum grade of 1 for your AA in order for this element to count towards completion of your module. The AA submission deadline is 31st March.

- **Accessing the assignments.** Assignments should be downloaded from the module page on the course website: http://dl.lshtm.ac.uk/programme/student/ep/student/modules/ep202.htm (see website section at the end of this document). Assignments are subject to change each year for all DL courses so you must ensure you download the assignments corresponding to the year in which you are submitting your assignment for marking. Assignments from previous years will not be accepted.

- Full details of how to submit assignments (using the on-line Assignment Management System) can be found in Chapter 8 of the Student Handbook.

- For all your assignment work, it is vital that you understand and apply principles of good academic writing, referencing and using source material, as well as avoiding plagiarism. Please refer to the Academic Writing Handbook for guidance on this – this can be found on the General Resources page of the student website http://dl.lshtm.ac.uk/programme/student/ep/student/general.htm.

Web conferencing

- The ‘WebBoard’ is our current web-based conference system* and an integral part of this course, putting you in touch with other students and with tutors on relevant modules. The WebBoard will be open from the first week of October onwards until the module exam.

- It is used for:
  - **Discussion of module content and queries.** We strongly recommend that you make use of the EP web conferencing system (WebBoard) to discuss issues relating to the course material. This gives an opportunity for you to ask questions and to take part in discussions initiated by fellow students. The conferences are monitored by tutors who will contribute to the discussions. There are conferences for different parts of the module, together with a General EPM202 conference, conferences for discussing FAs and AAs (AA conferences open in January), and, in May, conferences to help you prepare for the exams by discussing past exam questions. You may also email specific queries to
Module introduction

the Distance Learning Support Office (dlsupport@lshtm.ac.uk) who will refer your queries to one of the EPM202 tutors.

- **Messages from your Module Organiser and Distance Learning Support Office.** We use WebBoard as the primary means of communication of important messages between students and staff, and in early October, the Student Support Office will subscribe you to the mailing lists for the **EPM202 NoticeBoard (Students).** This conference will be ‘read-only’ with only Course Directors, Module Organisers and Student Support Office staff posting messages in these and it is essential that you read all messages posted there. Please log on regularly to WebBoard to view this module NoticeBoard, and check you are receiving these messages by email. Please do this at least once by the end of October at the latest. If you see NoticeBoard messages for this module on WebBoard that have not reached you by email, please contact the Distance Learning Support Office (dlsupport@lshtm.ac.uk)

  - **Accessing WebBoard.** Access is via the student website [http://dl.lshtm.ac.uk/](http://dl.lshtm.ac.uk/) (then click on EP). All students studying EP modules should have their own unique username and password to access the EP WebBoard. Please contact the Distance Learning Support Office (dlsupport@lshtm.ac.uk) if you do not have this. For information about using WebBoard in general, please see Chapter 7 of the Student Handbook.

*Please note that we may be upgrading our web conferencing software over the next few months. We will keep you informed of any changes.

**Website**

- Assignments and additional resources such as past exam questions, Frequently Asked Questions, list of module tutors, student evaluations etc can be downloaded from the **EPM202 module page** on the student website [http://dl.lshtm.ac.uk/programme/student/ep/student/modules/ep202.htm](http://dl.lshtm.ac.uk/programme/student/ep/student/modules/ep202.htm).

- There is also a **General Resources page** which includes documents such as the Student Handbook, list of course materials, and general exam guidance.

- **Accessing the student website.** You will be provided with the username and password to access the website at the beginning of the academic year. Please email dlsupport@lshtm.ac.uk if you do not know these.

Finally, a reminder that we are here to help! Do let us know if you have any queries at all – either by posting in the relevant WebBoard conference, or emailing dlsupport@lshtm.ac.uk.
Module Specification (Distance Learning)

In collaboration with University of London International Programmes

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><strong>Title:</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Statistical Methods in Epidemiology</strong></td>
</tr>
<tr>
<td>2.</td>
<td><strong>Module code:</strong></td>
</tr>
<tr>
<td></td>
<td>EPM202</td>
</tr>
<tr>
<td>3.</td>
<td><strong>Institution:</strong></td>
</tr>
<tr>
<td></td>
<td>Faculty of Epidemiology and Population Health</td>
</tr>
<tr>
<td></td>
<td>London School of Hygiene &amp; Tropical Medicine</td>
</tr>
<tr>
<td></td>
<td>Keppel Street</td>
</tr>
<tr>
<td></td>
<td>London</td>
</tr>
<tr>
<td></td>
<td>WC1E 7HT</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.lshtm.ac.uk/eph/">http://www.lshtm.ac.uk/eph/</a></td>
</tr>
<tr>
<td>4.</td>
<td><strong>Module Organisers:</strong></td>
</tr>
<tr>
<td></td>
<td>Jim Todd, Jo Haviland</td>
</tr>
<tr>
<td>5.</td>
<td><strong>Mode of study:</strong></td>
</tr>
<tr>
<td></td>
<td>Distance learning</td>
</tr>
<tr>
<td>6.</td>
<td><strong>Type:</strong></td>
</tr>
<tr>
<td></td>
<td>Elective (compulsory for some courses – see Section 18)</td>
</tr>
<tr>
<td>7.</td>
<td><strong>Duration and dates:</strong></td>
</tr>
<tr>
<td></td>
<td>Deadlines if taken as part of a formal award:</td>
</tr>
<tr>
<td></td>
<td>Application deadline:</td>
</tr>
<tr>
<td></td>
<td>30 June each year</td>
</tr>
<tr>
<td></td>
<td>Registration deadline:</td>
</tr>
<tr>
<td></td>
<td>31 August each year</td>
</tr>
<tr>
<td></td>
<td>Course registration duration:</td>
</tr>
<tr>
<td></td>
<td>Up to 5 years</td>
</tr>
<tr>
<td></td>
<td>Course starts:</td>
</tr>
<tr>
<td></td>
<td>1 October each year</td>
</tr>
<tr>
<td></td>
<td>Examination takes place:</td>
</tr>
<tr>
<td></td>
<td>Usually June each year (date to be confirmed)</td>
</tr>
<tr>
<td></td>
<td>Deadlines if taken as an individual module (i.e. not registered for formal award):</td>
</tr>
<tr>
<td></td>
<td>Application deadline:</td>
</tr>
<tr>
<td></td>
<td>31 August each year</td>
</tr>
<tr>
<td></td>
<td>Registration deadline:</td>
</tr>
<tr>
<td></td>
<td>30 November each year</td>
</tr>
<tr>
<td></td>
<td>Registration duration:</td>
</tr>
<tr>
<td></td>
<td>2 years</td>
</tr>
<tr>
<td></td>
<td>Module study starts:</td>
</tr>
<tr>
<td></td>
<td>1 October each year</td>
</tr>
<tr>
<td></td>
<td>Examination takes place:</td>
</tr>
<tr>
<td></td>
<td>Usually June each year (date to be confirmed)</td>
</tr>
<tr>
<td>8.</td>
<td><strong>Credit points:</strong></td>
</tr>
<tr>
<td></td>
<td>15 credit points will be awarded on successful completion of this module at Masters level (Level 7).</td>
</tr>
<tr>
<td>9.</td>
<td><strong>Notional Learning Hours (NLH):</strong></td>
</tr>
<tr>
<td></td>
<td>The module should take about 150 hours to complete. On average students will divide these learning hours as follows:</td>
</tr>
<tr>
<td></td>
<td>Directed self-study</td>
</tr>
<tr>
<td></td>
<td>70 hours</td>
</tr>
<tr>
<td></td>
<td>Self-directed learning</td>
</tr>
<tr>
<td></td>
<td>30 hours</td>
</tr>
<tr>
<td></td>
<td>Assessment, review and revision</td>
</tr>
<tr>
<td></td>
<td>50 hours</td>
</tr>
<tr>
<td>10.</td>
<td><strong>Aim:</strong></td>
</tr>
<tr>
<td></td>
<td>This module aims to provide students with the key statistical knowledge and skills needed to analyse and interpret data from the common forms of epidemiological studies.</td>
</tr>
<tr>
<td>11.</td>
<td><strong>Learning objectives:</strong></td>
</tr>
<tr>
<td></td>
<td>On completion of this module students should be able to:</td>
</tr>
<tr>
<td></td>
<td>• explain the basic statistical measures and concepts underlying the analyses of epidemiological data</td>
</tr>
<tr>
<td></td>
<td>• be familiar with a comprehensive set of statistical methods suitable for a wide range of epidemiological situations</td>
</tr>
<tr>
<td></td>
<td>• select appropriate statistical techniques for the analysis of data from epidemiological studies</td>
</tr>
<tr>
<td></td>
<td>• identify specific issues relevant to case-control and cohort studies</td>
</tr>
<tr>
<td></td>
<td>• demonstrate an understanding of statistical modelling techniques</td>
</tr>
<tr>
<td></td>
<td>• investigate confounding and interaction in epidemiological data using both stratified analyses and statistical modelling methods</td>
</tr>
<tr>
<td></td>
<td>• interpret the results of statistical procedures and draw appropriate conclusions.</td>
</tr>
</tbody>
</table>
12. Content: Module content is structured around the self-study sessions listed below:

<table>
<thead>
<tr>
<th>Module</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM01</td>
<td>Introduction/Measures of effect</td>
</tr>
<tr>
<td>SM02</td>
<td>Cohort studies</td>
</tr>
<tr>
<td>SM03</td>
<td>Survival analysis</td>
</tr>
<tr>
<td>SM04</td>
<td>Case-control studies</td>
</tr>
<tr>
<td>SM05</td>
<td>Likelihood</td>
</tr>
<tr>
<td>SM06</td>
<td>Multivariable analysis</td>
</tr>
<tr>
<td>SM07</td>
<td>Logistic regression 1</td>
</tr>
<tr>
<td>SM08</td>
<td>Logistic regression 2</td>
</tr>
<tr>
<td>SM09</td>
<td>Logistic regression 3</td>
</tr>
<tr>
<td>SM10</td>
<td>Matched case-control studies</td>
</tr>
<tr>
<td>SM11</td>
<td>Introduction to Poisson and Cox regression</td>
</tr>
<tr>
<td>SM12</td>
<td>Strategies of analysis</td>
</tr>
<tr>
<td>SM13</td>
<td>Summary.</td>
</tr>
</tbody>
</table>

In the first half of the module the focus is on issues specific to different types of study. The second half of the module deals with statistical modelling and multivariable analyses. The combined materials will enable students to choose and use the techniques appropriate for estimation and hypothesis testing in selected situations.

13. Learning methods: Learning is self-directed against a detailed set of learning objectives using the materials provided. The key learning methods are:
- Reading and reflecting on CAL (computer-assisted learning) materials which introduce, explain and apply the principles and methods covered in the module.
- Reading and reflecting on paper-based materials which support the learning in the CAL sessions.
- Completing paper and computer-based practical exercises.
- Accessing academic support which is available from the module tutors through the web-based discussion forum in which students are encouraged to participate.
- Completing the formative assignment and reflecting on written feedback from module tutors.
- Completing the assessed assignment and reflecting on written feedback from module tutors.

14. Study resources provided:

CD-Rom - EPM201/202  
EPM202 Statistical Methods in Epidemiology Workbook & Reader.

Software: Stata

Textbook: Essential Medical Statistics (Kirkwood, Sterne).

Registered students have access to the School's online library resources. Students who are taking this as an individual module or as part of the MSc/PG Diploma (CF) PH course will also have online access to the MSc EP core electronic study materials (this access will exclude tutor support and associated textbooks).
15. Assessment procedures: Formal assessment of the module will consist of one Assessed Assignment (comprising 30% of the total grade for the module). Students are also assessed by a two-hour written examination (70% of the total grade for the module).

Examinations are normally held in a student's country of residence, in one of over 650 examination centres worldwide. They are arranged mainly through Ministries of Education or the British Council. A local fee will be payable. A list of examination centres can be found at [http://www.londoninternational.ac.uk/current_students/general_resources/exams/exam_centres/index.shtml](http://www.londoninternational.ac.uk/current_students/general_resources/exams/exam_centres/index.shtml).

If students fail an examination at the first entry they will be allowed one further attempt, the following year.

16. Prerequisites: Students should have completed EPM101, EPM102, EPM103 and EPM105 (core modules) or have equivalent experience.

MSc Public Health students, and PG Diploma Public Health students following the credit framework structure, who wish to take this module are required to have obtained a pass grade in the PHM102 Basic Statistics for Public Health core module: a grade of at least 4 is recommended. In particular, for adequate preparation, students should have studied the optional PHM102 CD-ROM session 14 and carried out all the Stata exercises in PHM102.

Those wishing to study this module must have regular access to the internet to benefit from library facilities, participate in web-based conference discussions and submit assignments.

Students must meet the standard of English required to study this course. See [http://www.lshtm.ac.uk/prospectus/english.html](http://www.lshtm.ac.uk/prospectus/english.html).

17. Attendance: No maximum number

18. Selection, if applicable: This module is available to those registered for the MSc Epidemiology or Public Health courses; alternatively, it can be taken as an Individual Module. It is a compulsory module for those studying for the PG Diploma Epidemiology course under the credit framework scheme, and for those studying the MSc Epidemiology course. It can also be taken by those studying for the PG Diploma Public Health course under the credit framework structure.


20. Scholarships: None available

21. External accreditation: None

22. Application process: Applications are managed by the University of London International Programmes (website: [http://www.londoninternational.ac.uk/](http://www.londoninternational.ac.uk/)).

23. Further enquiries: Enquiries may be emailed to distance@lshtm.ac.uk.
Section 1

Stata glossary

How to use this glossary

This glossary provides a listing of all Stata commands which are useful in statistical analysis, including those used in the practical sessions. It is intended as a quick reference, primarily to remind you of command names. Commands are given under the headings listed below, and may appear in more than one place.

- utilities
- data manipulation and management
- descriptive statistics
- general statistics
- cohort/survival analysis
- case-control/cross-sectional analysis
- regression models.

Once you have found the command you need, you can type help followed by the command name to get further information. The glossary is not intended to give a full description of the use of each command. This is given in the official Stata manuals.

Utilities

clear clear data from memory
display display values from functions
do execute commands from a file
exit exit Stata
help obtain online help
log record all commands and output entered during session into a file
lookup obtain online help; useful if the term you use is not a command term
save save data into a Stata file
use use a Stata dataset
**Data manipulation and management**

- **codebook** to obtain information on variable type, label format and summary
- **collapse** collapse data into a table
- **count** count number of observations
- **describe** (or F3) describe contents of dataset in memory
- **drop** eliminate variables
- **encode** create a numeric variable from a string variable
- **expand** duplicate observations
- **format** specify permanent display format eg number of decimal places to display or how date will appear
- **generate** create new variable
- **infile** read non-Stata data into memory eg ASCII text file
- **input** enter data from keyboard
- **keep** keep a subset of variables eg for new dataset
- **label** manipulation of labels for variables, categories or datasets
- **list** list values of variables
- **merge** merge two sorted datasets
- **mvdecode** change a numeric code for a missing value to the Stata-defined to missing value (.)
- **mvencode** change missing (.) to a coded value
- **outfile** write data to a non-Stata format
- **quietly** carry out next command without giving output
- **recode** recode numeric to categorical variables
- **rename** change name of existing variable
- **replace** change value of an existing variable in all or specified subsets of records
- **set** set general options in Stata
- **sort** sort records according to a variable

**Descriptive statistics**

- **centile** calculates (per) centiles of given variable(s)
- **list** list values of variables
- **graph** display observations graphically
- **hist** histogram of categorical variable
- **summarize** display summary statistics
**General statistics**

- **anova**  analysis of variance
- **correlate**  correlation between variables
- **oneway**  one-way analysis of variance
- **ranksum**  Wilcoxon ranksum test
- **table**  multiple two-way tables of frequencies and summary statistics
- **tabulate**  one- and two-way tables of frequencies
- **ttest**  mean comparison test for small samples
- **ztest**  mean comparison test

**Cohort/survival analysis**

- **poisson**  Poisson regression (output on log scale)
- **stmh**  Mantel-Haenszel rate ratios
- **strate**  table of rates
- **stset**  setting the survival time variables
- **stcox**  Cox regression (output on log scale)
- **stsplit**  expand data according to a Lexis diagram
- **sts list**  listing of Kaplan-Meier survival functions
- **sts graph**  graph of Kaplan-Meier survival functions
- **sts graph, na**  Nelson-Aalen cumulative hazards
- **sts test**  Logrank test

**Case-control/cross-sectional analysis**

- **clogit**  conditional logistic regression (output on log scale)
- **glm**  generalised linear models
- **logit**  logistic regression (output on log scale)
- **logistic**  logistic regression (output on antilog scale)
- **match**  table of matched data
- **mhodds**  Mantel-Haenszel odds ratios
- **tabodds**  table of odds (for case-control studies the ‘odds’ shown are not real odds but a constant multiple of them)
Regression models

clogit  conditional logistic regression (output on log scale)
glm    generalised linear models
logit   logistic regression (output on log scale)
logistic logitistic regression (output on antilog scale)
lrtest  likelihood ratio test
poisson Poisson regression (output on log scale)
regress linear regression
xi: command i.varnames allows to fit a regression model where categorical explanatory variables are included as a set of indicators.
Section 2
Description of datasets
Section 2

Description of datasets

Dataset: Whitehall

Cohort study of risk factors for mortality in an occupational cohort

Data on risk factors for ischaemic heart disease (IHD) were collected between 1967-69 for a total of 19,183 male civil servants from various departments around Whitehall (London). The data were collected by self-administered questionnaire and a screening examination. Survey participants were identified and flagged at the National Health Service Central Registry and a coded copy of the death certificate provided for each subsequent death. Whitehall.dta refers to a 10% sample of the survey participants.

The coding of this dataset is given on a Stata help file.

Type help whitehall in Stata for this information.

Dataset: Mwanza

Case-control study of risk factors for HIV infection among women, Mwanza, Tanzania

As part of a prospective study of the impact of STD control on the incidence of HIV infection in Mwanza, Tanzania, a baseline survey of HIV prevalence was carried out in 12 communities. All seropositive women (15 years and above) were revisited and, where possible, interviewed about potential risk factors for HIV infection using a standard questionnaire. Data were collected on the following: personal information (age, education, religion, ethnicity, marital status, etc); residence and travel history; non-sexual risk factors (blood transfusions, injections, etc); sexual behaviour (number of regular partners, casual partners, etc); condom use and history of STDS; AIDS/STD risk perception. In addition to interviewing HIV +ve women, a random sample of HIV -ve women were selected from the population lists prepared during the baseline survey and these women were also revisited and, where possible, interviewed. No matching of controls with cases was performed.

A total of 189 cases and 574 controls were recruited, follow-up rates of 67% and 74% respectively.

The coding of this dataset is given on a Stata help file.

Type help mwanza in Stata for this information.
Section 2: Description of datasets

Dataset: Oncho

Prevalence study of onchocerciasis in Sierra Leone

Onchocerciasis (commonly known as River Blindness) is a chronic filarial disease found in sub-Saharan Africa and some parts of Central and South America. Adult worms of *Onchocerca volvulus* are found in nodules, mainly around the pelvic girdle. Microfilariae (mff) discharged by the female worm migrate through the skin, often causing intense rash and itching, dyspigmentation and atrophy. The mff also often migrate to the eye and cause visual impairment and ultimately blindness.

Transmission is via the bite of infected female blackflies of *Simulium* species. Mff, ingested by a blackfly while feeding on an infected person, penetrate the thoracic muscles of the fly, develop into infective larvae and enter a bite wound during a subsequent blood meal.

An onchocerciasis project supported by the British Medical Research Council was set up in 1982 in the Bo district of Sierra Leone. The aims of the project were to study epidemiological, clinical, immunological and entomological aspects of the disease. Prevalence surveys were undertaken in villages selected on the basis of potential high endemicity, being situated on or near rivers which are the breeding sites for the *Simulium damnosum* blackfly. Of the twelve villages included in the present dataset, five were situated in the south and east of the country in the 'forest' zone (secondary forest or oil palm bush) and the other seven were in the 'savannah' zone (woodland savanna/forest-savanna mosaic) in the north and northeast of the country.

A census was taken of each village, and all villagers over the age of five years were asked to participate in the study. Coverage was over 90% in all but one of the selected villages. Diagnosis was made by taking a skin-snip which was placed in saline and allowed to dry, and then counting, under the low power of a compound microscope, the mff which had emerged. Both a clinical and an ocular examination were also performed. The latter was conducted in a darkened room, using a slit-lamp, and the presence of eye lesions and of mff in the cornea or anterior chamber were recorded.

The coding of this dataset is given on a Stata help file.

Type `help oncho` in Stata for this information.

Dataset: Trinidad

All males aged 35-74 years who were living in two neighbouring suburbs of Port of Spain, Trinidad, in March 1977 were eligible and entered into the study. Baseline data were recorded for 1,343 men on a range of risk factors including ethnic group, blood pressure, glucose and lipoprotein concentrations, diabetes mellitus, and cigarette and alcohol consumption. All subjects were then visited annually at home, and morbidity and mortality records were compiled. Regular inspection of hospital records, death registers and obituaries were also used to update the records. Those who had moved away (or abroad) were contacted annually by postal questionnaire.
Section 2: Description of datasets

and were also seen if they returned to Port of Spain. By these means, loss to follow-up was kept very low. Follow-up of the study cohort finished at the end of 1986, giving a study period of almost ten years.

The file trinidad contains data on selected risk factors for the subset of men aged 60 years or over. There were 318 men in this group, and 88 deaths were recorded. Of these deaths, 22 were attributed to cardiovascular disease.

The coding of this dataset is given on a Stata help file.

Type help trinidad in Stata for this information.

Dataset: Diet

These data arose from a study of the use of a weighed diet assessment over 7 days in epidemiological studies. Men who had their dietary intake measured in this way were followed up for various outcomes. The data in diet.dta relate subsequent incidence of coronary heart disease (CHD) to dietary energy intake.

The coding of this dataset is given on a Stata help file.

Type help diet in Stata for this information.

Dataset: Chilumba

This file chilumba.dta contains data from an unmatched case-control study of risk factors for leprosy. Between 1980 and 1984 a population of approximately 112,000 people living in Karonga District, Northern Malawi, were screened for leprosy. Individuals found to have leprosy were not followed further. The remaining population was followed until 1989. During the follow-up period 252 new cases of leprosy were identified. 1008 controls without leprosy at baseline were selected at random from the screened population. Whether an individual had been vaccinated with BCG was assessed by examining whether they possessed a typical BCG scar when screened. BCG was introduced into Karonga District mainly in mass vaccination campaigns in schools during the late 1970s.

The coding of this dataset is given on a Stata help file.

Type help chilumba in Stata for this information.
Section 3
Practical sessions
(1-4 and 7-10 only)


Practical 1

Measures of effect

Aim

The aim of this practical is to introduce the Stata commands used to calculate the main measures of disease and effect used in epidemiology: risks, risk ratios, rates, rate ratios, and odds ratios.

Objectives

By the end of this session you will be able to:

- obtain risks and risk ratios
- define follow-up information for a cohort study using Stata
- obtain rates and rate ratios for a cohort study
- obtain odds ratio for a case-control study.

Planning your session

To complete the work in this session:

- You should have worked through Session 1 of the Statistical Methods in Epidemiology CAL material
- You should have copied all the datasets and files to your working directory for these practical exercises (e.g. saved to a directory called sme). The specific datasets for this session are whitehal.dta and mwanza.dta
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created (e.g. using Windows Explorer) for your workbook practicals:

e.g. cd c:\sme

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

e.g. log using smeprac1.log

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the cd command.
**Remember:**

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.

**Note:**

All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0 or older, and are compatible with later versions of Stata.

---

**Reading in the data and initial examination**

For the first part of this practical session you will use data from the Whitehall study (*whitehal.dta*).

To read in the dataset type:

```
use whitehal
```

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

```
help whitehal
```

You are going to examine the effect of smoking on mortality from any cause.

- **all** is the variable name for all cause mortality coded: 1=death from any cause; 0=alive
- **smok** is the variable name for smoking category coded: 1=never smoked; 2=ex-smoker; 3=1-14 cigarettes per day; 4=15-24 cigarettes per day; 5=25+ cigarettes per day

To find out how many deaths from all causes there were, type:

```
tab all
```

```
   death from all causes | Freq. | Percent | Cum.     
----------------------|-------|---------|----------
        0              | 1274  | 75.97   | 75.97    
        1              | 403   | 24.03   | 100.00   
----------------------|-------|---------|----------
Total                | 1677  | 100.00  |          
```
There were 403 deaths from all causes (24.03%).
Similarly to see the distribution of the cohort for smoking categories, type:

```
tab smok
```

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>317</td>
<td>18.90</td>
<td>18.90</td>
</tr>
<tr>
<td>2</td>
<td>646</td>
<td>38.52</td>
<td>57.42</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>18.49</td>
<td>75.91</td>
</tr>
<tr>
<td>4</td>
<td>279</td>
<td>16.64</td>
<td>92.55</td>
</tr>
<tr>
<td>5</td>
<td>125</td>
<td>7.45</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 1677 | 100.00 |

So, there were 317 men who had never smoked, 646 ex-smokers, etc.

To split the cohort into current smokers and non-smokers you must create a new variable that takes the values 1 for never or ex-smokers and 2 for smokers. To do this use the following commands:

```
gen smok2 = smok
recode smok2 1/2=0 3/5=1
```

It is good practice always to check that the new variable has been coded correctly. You can do this by tabulating it against the old one. Type:

```
tab smok smok2
```

<table>
<thead>
<tr>
<th>never/ex/1- 14/15-24/25 +</th>
<th>smok2</th>
<th>never/ex/1- 14/15-24/25 +</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>310</td>
<td>3</td>
<td>310</td>
</tr>
<tr>
<td>4</td>
<td>279</td>
<td>4</td>
<td>279</td>
</tr>
<tr>
<td>5</td>
<td>125</td>
<td>5</td>
<td>125</td>
</tr>
</tbody>
</table>

Total | 963 | 714 | 1677 |

**So those who never smoked and the ex-smokers are coded as 0 in smok2 and anyone who smokes at least one cigarette is coded as 1.**

**Calculation of risks and risk ratios**

There are 3 ways we can obtain risks, the third also gives risk ratios automatically.

- To examine the risk of death according to smoking status you can simply use the `tab` command. It is easy to obtain risks using this command. Type:

```
tab smok2 all
```
To add row percentages, which gives the risk, type:

```
tab smok2 all, row
```

<table>
<thead>
<tr>
<th>death from all causes</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>smok2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>0</td>
<td>795</td>
<td>168</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td>82.55</td>
<td>17.45</td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td>479</td>
<td>235</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td>67.09</td>
<td>32.91</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>1274</td>
<td>403</td>
<td>1677</td>
</tr>
<tr>
<td></td>
<td>75.97</td>
<td>24.03</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The row percent for all = 1 gives the risk of death in each group multiplied by 100.

Another way to derive the risk of death is to use the **summarize** option with the `tab` command.

```
tab smok2, summarize(all)
```

<table>
<thead>
<tr>
<th>Summary of death from all causes</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>smok2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>0</td>
<td>.17445483</td>
<td>.37969731</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td>.1745</td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.32913165</td>
<td>.47022728</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td>0.329</td>
<td>0.329</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>.24031008</td>
<td>.42739919</td>
<td>1677</td>
</tr>
<tr>
<td></td>
<td>0.2403 (17.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Because **all** is coded 0/1, its mean is the total number of deaths divided by the total number of observations (i.e. the risk).

Mean = \( \sum 0 + 1 + 0 + \ldots / \) \( N \) = total number of deaths / total number of observations.

We can also produce a table of risk measures and effect measures using an Epitab command specifically designed for risks. Type:

```
cs all smok2
```

<table>
<thead>
<tr>
<th>smok2</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>235</td>
<td>168</td>
<td>403</td>
</tr>
<tr>
<td>Cases</td>
<td>479</td>
<td>795</td>
<td>1274</td>
</tr>
<tr>
<td>Noncases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>714</td>
<td>963</td>
<td>1677</td>
</tr>
<tr>
<td>Risk</td>
<td>.3291317</td>
<td>.1744548</td>
<td>.2403101</td>
</tr>
</tbody>
</table>

| Risk difference | .1546768 | .112695   | .1966586 |
| Risk ratio      | 1.88663  | 1.587309  | 2.242394 |
| Attr. frac. ex. | .4699543 | .3700029  | .554048  |
| Attr. frac. pop | .2740428 |           |         |
| Point estimate  |         | [95% Conf. Interval] |
| chisq(1) = 53.73 |  Pr>chisq = 0.0000 |
The risks are 0.329 and 0.174 for smokers and non-smokers respectively and we also have the risk ratio = 1.89 (95% CI 1.59 to 2.24)

The output gives a two-by-two table with the same cells as the `tab` command, and below that the risk estimates. Then it gives the risk difference (i.e. risk among exposed minus risk among unexposed), followed by the risk ratio. The final two parameters apply if smoking really is causal. They are the percentage of outcomes among the exposed attributable to the exposure (current smoking) and the percentage of outcomes among the population attributable to the exposure, assuming that your sample accurately represents the percentage of smokers in the population.

### Defining follow-up information

Before you are able to use most of the Stata commands appropriate for the analysis of follow-up data, you must first define the dates of entry and exit into the study and the outcome (or ‘failure’) variable. This is done with the command `stset`. This command takes the general form below (do not be put off! take each item one at a time).

```stata
stset timeout, fail(fail) id(idno) origin(start) enter(timein) scale(number)
```

- "timeout" and the words in brackets are replaced by the names of the relevant variables in your dataset.
- The `timeout` variable defines the date of exit.
- In the Whitehall dataset the variable is actually called `timeout`.
- The `fail` variable is the outcome variable (or reason for exit in survival data).
- In the Whitehall dataset you are examining the outcome all-cause mortality using the variable `all`.
- The `idno` variable contains the subject identity number (`id` in this dataset).
- The `start` variable defines the date the subject becomes at risk (e.g. the date someone starts a job in the nuclear industry or the date they were born, depending on the exposure).
- The `timein` variable defines the date of entry into the study. The date of entry for some studies is the same as the origin time, e.g. in clinical trials. If `enter` is not specified, Stata assumes that the entry and origin times are the same.
- This variable is called `timein` in the Whitehall dataset.
- The value of `number` in the scale specifies the time units for analyses. Since dates in Stata are expressed in days since 1 Jan 1960, number is often set at 365.25 to convert the time units to person-years.

For this exercise, for the minimum specification type:

```stata
stset timeout, fail(all) id(id) origin(timein) scale(365.25)
```
### Calculation of rates

Note: The commands that start with `st` are used with follow-up data. `st` = survival time. All the `st` commands will only work if the `stset` command has been used.

Once the `stset` command has been used, it is possible to calculate rates using the command `strate`.

To obtain all-cause mortality rates for smokers and non-smokers, type:

```stata
strate smok2
```

<table>
<thead>
<tr>
<th>smok2</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>168</td>
<td>1.6e+04</td>
<td>0.0102673</td>
<td>0.0088264</td>
<td>0.0119434</td>
</tr>
<tr>
<td>1</td>
<td>235</td>
<td>1.1e+04</td>
<td>0.0209024</td>
<td>0.0183937</td>
<td>0.0237532</td>
</tr>
</tbody>
</table>

In the output,

D gives number of deaths

Y the number of person-years from `timein` to `timeout`

The rate is per year (D/Y).

The final columns show 95% confidence intervals.

The rate for the combined group of non-smokers and ex-smokers is 0.010 per person-year (168/1.6x10^4). The rate for smokers is 0.021 per person-year (235/1.1x10^4).
To obtain the rate per 1000 person-years, type:

```
strate smok2, per(1000)
```

```
failure _d: all
analysis time _t: (timeout-origin)/365.25
origin: time timein
id: id
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)

<table>
<thead>
<tr>
<th>smok2</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>168</td>
<td>16.3626</td>
<td>10.2673</td>
<td>8.8264</td>
<td>11.9434</td>
</tr>
<tr>
<td>1</td>
<td>235</td>
<td>11.2427</td>
<td>20.9024</td>
<td>18.3937</td>
<td>23.7532</td>
</tr>
</tbody>
</table>

To obtain all-cause mortality rates for the finer categories of smoking using the original variable for smoking, type:

```
strate smok, per(1000)
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)

<table>
<thead>
<tr>
<th>smok</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33</td>
<td>5.5921</td>
<td>5.9012</td>
<td>4.1953</td>
<td>8.3007</td>
</tr>
<tr>
<td>2</td>
<td>135</td>
<td>10.7706</td>
<td>12.5342</td>
<td>10.5885</td>
<td>14.8373</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>5.0148</td>
<td>17.7473</td>
<td>14.4180</td>
<td>21.8454</td>
</tr>
<tr>
<td>4</td>
<td>98</td>
<td>4.2911</td>
<td>22.8382</td>
<td>18.7360</td>
<td>27.8385</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>1.9368</td>
<td>24.7826</td>
<td>18.6761</td>
<td>32.8857</td>
</tr>
</tbody>
</table>

Death rates increase in ex-smokers compared to non-smokers, and with increasing quantities smoked.

You can plot a graph of deaths by smoking status:

```
strate smok, graph per(1000) ylab(0(10)30, angle(horizontal))
```
The graph option `ylab(0(10)30, angle(horizontal))` specifies that the y-axis should be labelled between 0 and 30 in units of 10, with the angle of text set to horizontal (default is vertical).

### Calculation of rate ratios

Rate ratios are derived using the command `stmh`.

To compare death rate among smokers with death rate among non-smokers, type:

```bash
stmh smok2
```

```bash
failure _d:  all
analysis time _t:  timeout/365.25
enter on or after:  time timein
id:  id
```

Maximum likelihood estimate of the rate ratio comparing `smok2==1` vs. `smok2==0`

RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.036</td>
<td>51.63</td>
<td>0.0000</td>
<td>1.670</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.482</td>
</tr>
</tbody>
</table>

From this we can conclude that smokers have twice the death rate (2.04) of non-smokers (95% CI 1.67 to 2.48). Note that the category with the lower code value (code 0 = never/ex-smokers) is taken as the reference category, unless you specify otherwise using the `compare` option `c(...)`.

The output shows the rate ratio for current smokers compared to never/ex-smokers, together with confidence limits. Using the output from the `strate smok2` command, check that the rate ratio is the rate for smokers divided by the rate for non-smokers.

\[
20.90 / 10.27 = 2.04
\]
Note that the rate ratio is greater than the risk ratio (1.89) that you calculated on page 1.5. This is because there was less person-time among smokers than non-smokers.

Finally... for the cohort analysis.

If you want to save the variables you have created, save the dataset but make sure you use a different name. This ensures you keep a clean dataset available. Type:

```
save whitehall2
```

### Calculation of odds ratios

For this part of the practical you will use data from the case-control study of HIV infection in women in Mwanza, Tanzania. We will examine this dataset in more detail in Practical 4.

```
use mwanza, clear
```

For a description of the study and variables in the dataset, type:

```
help mwanza
```

You will examine whether there is an association between HIV infection and duration of education.

**case** is the variable name for HIV infection
coded: 1=case; 0=control

**ed** is the variable name for level of education
coded: 1=no formal education (none/adult only); 2=1-3 years;
3=4-6 years; 4=7+ years

To find out how many cases and controls there were, type:

```
tab case
```

<table>
<thead>
<tr>
<th>Case/control</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>574</td>
<td>75.23</td>
<td>75.23</td>
</tr>
<tr>
<td>1</td>
<td>189</td>
<td>24.77</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>763</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Similarly to see the distribution of women by level of education, type:

```
tab ed
```

<table>
<thead>
<tr>
<th>Education</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>312</td>
<td>40.89</td>
<td>40.89</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>9.83</td>
<td>50.72</td>
</tr>
<tr>
<td>3</td>
<td>365</td>
<td>47.84</td>
<td>98.56</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>1.44</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>763</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

There were 189 cases and 574 controls.

40.9% of the women had no formal education.
To examine the difference in education levels in cases and controls, type:

```
tab case ed, row chi
```

<table>
<thead>
<tr>
<th>Case/control</th>
<th>Education</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>263</td>
<td>51</td>
<td>255</td>
<td>5</td>
<td>574</td>
</tr>
<tr>
<td></td>
<td>45.82</td>
<td>8.89</td>
<td>44.43</td>
<td>0.87</td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td>49</td>
<td>24</td>
<td>110</td>
<td>6</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>25.93</td>
<td>12.70</td>
<td>58.20</td>
<td>3.17</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>312</td>
<td>75</td>
<td>365</td>
<td>11</td>
<td>763</td>
</tr>
<tr>
<td></td>
<td>40.89</td>
<td>9.83</td>
<td>47.84</td>
<td>1.44</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Pearson chi2(3) = 26.7371  Pr = 0.000

The `chi` option tests the null hypothesis that the two distributions of education level are the same.

The null hypothesis is rejected (P<0.001). Those with no education seem to be under-represented among cases relative to controls (26% against 46%). Also note that the group with the highest education (code 4) is small, therefore any estimates of effect for this group will have wide confidence intervals.

Because this is a case-control study, and the cases and controls therefore have different probabilities of selection, we cannot calculate the odds of disease for each education group from the above table. However, we can calculate odds ratios. This will be further explained in Session 4 of the CAL material.

Create a new variable `ed2` which takes the value 1 for women with no formal education and value 2 for those with some education (codes 2-4 of `ed`).

```
   gen ed2 = ed
   recode ed2 3/4=2
```

To check that the new variable has been coded correctly, tabulate it against the original variable:

```
tab ed ed2
```

<table>
<thead>
<tr>
<th>Education</th>
<th>ed2</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>312</td>
<td>0</td>
<td>312</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>75</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>365</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>312</td>
<td>451</td>
<td>763</td>
<td></td>
</tr>
</tbody>
</table>

Note: `stset` commands do not apply for case-control data (or cross-sectional data).

To derive the odds ratio we use the `mhodds` command:

```
mhodds case ed2
```
Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs ed2==1

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.416169</td>
<td>23.25</td>
<td>0.0000</td>
<td>1.668360 3.499168</td>
</tr>
</tbody>
</table>

It is important to take account of missing values when analysing data. The variable skin records skin incisions or tattoos (coded: 1=no, 2=yes, 9=missing).

To tabulate skin and obtain odds ratio estimates, type:

tag skin

<table>
<thead>
<tr>
<th>Skin</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>incisions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>or tattoos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>342</td>
<td>44.82</td>
<td>44.82</td>
</tr>
<tr>
<td>2</td>
<td>420</td>
<td>55.05</td>
<td>99.87</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.13</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>763</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

mhodds case skin

Score test for trend of odds with skin

(The OR estimate is an approximation to the odds ratio for a one unit increase in skin)

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.191688</td>
<td>1.40</td>
<td>0.2368</td>
<td>0.891231 1.593439</td>
</tr>
</tbody>
</table>

mhodds case skin, c(2,1)

Maximum likelihood estimate of the odds ratio
Comparing skin==2 vs skin==1

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.326275</td>
<td>2.74</td>
<td>0.0977</td>
<td>0.948473 1.854564</td>
</tr>
</tbody>
</table>

The option c(2,1) is telling Stata to make code 1 the reference group and gives us the relative odds for code 2 compared with code 1.

Note: without this option Stata will fit a trend through 1, 2 and 9, which is clearly not desirable, since 9 encodes missing data.
Now set the missing value for **skin** to system-missing (.).

```
recode skin 9 =.
```

and again type:

```
mhodds case skin
```

This gives the same output as using the `c(2,1)` option.

Note: It is good practice to specify the `compare` option to make sure that Stata makes the correct comparison.

Remember to close your log file at the end of your session.

```
log close
```

---

**Review exercise (time for you to have a go... )**

Now use Stata to answer the following questions for the Whitehall and the Mwanza datasets. Solutions to these exercises are in Section 4 of this Workbook.

Remember to read in each dataset with the `use` command, and save and clear your dataset before you read in a new file.

Read in the Whitehall dataset you saved earlier in the session. The dataset that contains the new recoded variables is called `whitehall2.dta`.

1. How many men in the dataset died of CHD?
2. What is the risk of death from CHD in each smoking category of `smok2`?
3. What is the risk ratio of current smokers to non-smokers?
4. Use the `stset` command to define follow-up details for the outcome CHD.
5. Use `strate` to obtain the CHD mortality rates for current smokers and non-smokers.
6. What is the rate of deaths due to CHD per 1000 person-years for current smokers and non-smokers?
7. Use `stmh` to obtain the rate ratio of CHD death rates for smokers to non-smokers.
8. Read in the Mwanza data.
9. How many women had ever used a condom (the variable name for condom use is `usedc`)?
10. What proportion of cases had ever used a condom?
11. Use `mhodds` to obtain the odds ratio for condom use and HIV infection. (Remember to account for the missing values for `usedc`
Cohort studies

Aim

The aim of this practical session is to learn how to compute Mantel-Haenszel summary rate ratios (RRs) and decide whether a summary measure is appropriate.

Objectives

By the end of this session you will be able to:

- define the outcome and time variables of the study
- obtain rates and rate ratios
- produce stratum-specific rate ratio estimates for the exposure of interest according to levels of the potential confounder
- assess whether the stratifying variable confounds or modifies the effect of the exposure of interest

Planning your session

To complete the work in this session:

- You should have worked through Session 2 of Statistical Methods in Epidemiology.
- The specific dataset for this session is whitehal.dta.
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

e.g. cd c:\sme

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

e.g. log using smeprac2.log

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the cd command.

Remember:

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.
Note:
All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0 or older, and are compatible with later versions of Stata.

Reading in the dataset and identifying relevant variables

Throughout this session you will be analysing the Whitehall dataset.
To read in the dataset, type:

```
use whitehal
```

You will see all the variables for the dataset listed in the Variables window.
For a description of the study and variables in the dataset, type:

```
help whitehal
```

You are going to examine the effect of grade of employment on mortality from any cause, adjusted for age at entry to the study.

- **all** is the variable name for all cause mortality
coded: 1=death from any cause; 0=otherwise
- **grade** is one of the variables for grade of work
coded: 1=admin & professional/executive; 2=clerical & other
- **agein** age at entry in years

Defining follow-up information

Remember from Practical 1, before you are able to analyse follow-up data you must first define the dates of entry and exit into the study and the outcome (or ‘failure’) variable. This is done using the `stset` command.

The outcome is the overall mortality **all**, whereas the time variables are **timein** and **timeout**. Both of these are expressed in days, so we need to set the scale to be 365.25 days to produce analyses in terms of person-years. Type:

```
stset timeout, fail(all) origin(timein) id(id) scale(365.25)
```

```
id: id
failure event: all != 0 & all <.
obsv. time interval: [timeout[_n-1], timeout]
exit on or before: failure
t for analysis: (time-origin)/365.25
origin: time timein
```

```
1677  total obs. 0  exclusions
1677  obs. remaining, representing 1677  subjects 403  failures in single failure-per-subject data 27605.37  total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 19.38123
```
There are 1677 subjects in the dataset. The follow-up time starts at entry, hence the minimum is zero, and lasts up to the maximum of 19.4 years.

**Stratum-specific rates**

To investigate how overall mortality varies according to age at entry to the study, you need to recode age at entry into suitable groups, then use the command `strate` to obtain rates within each age strata.

The variable `agein` holds age at entry into the study. To categorise it into groups (40-44, 45-49, 50-54, ..., 65-69) you can use the command `egen` with the cut option.

Note: Adding the option `label` defines labels for each category of `agecat`. Type:

```
egen agecat=cut(agein), at(40,45,50,55,60,65,70) label
```

You can use the `tab` command to check the distribution of `agecat`, type:

```
tab agecat, nolabel
```

<table>
<thead>
<tr>
<th>agecat</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>277</td>
<td>16.52</td>
<td>16.52</td>
</tr>
<tr>
<td>1</td>
<td>445</td>
<td>26.54</td>
<td>43.05</td>
</tr>
<tr>
<td>2</td>
<td>362</td>
<td>21.59</td>
<td>64.64</td>
</tr>
<tr>
<td>3</td>
<td>340</td>
<td>20.27</td>
<td>84.91</td>
</tr>
<tr>
<td>4</td>
<td>215</td>
<td>12.82</td>
<td>97.73</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>2.27</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>1677</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Now, to examine how mortality rates change with age at entry you can use `strate`. Type:

```
strate agecat, per(1000)
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals (1677 records included in the analysis)

<table>
<thead>
<tr>
<th>agecat</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-</td>
<td>24</td>
<td>4.9186</td>
<td>4.8794</td>
<td>3.2705</td>
<td>7.2798</td>
</tr>
<tr>
<td>45-</td>
<td>45</td>
<td>7.8549</td>
<td>5.7289</td>
<td>4.2714</td>
<td>7.6729</td>
</tr>
<tr>
<td>55-</td>
<td>118</td>
<td>5.2820</td>
<td>22.3400</td>
<td>18.6519</td>
<td>26.7573</td>
</tr>
<tr>
<td>60-</td>
<td>101</td>
<td>3.0948</td>
<td>32.6349</td>
<td>26.8525</td>
<td>39.6625</td>
</tr>
<tr>
<td>65-</td>
<td>33</td>
<td>0.3952</td>
<td>83.4921</td>
<td>59.3567</td>
<td>117.4412</td>
</tr>
</tbody>
</table>

The rates increase quite dramatically with each age category. This is best displayed in a graph. To produce a graph displaying the mortality trend with age at entry, use the `graph` option with `strate`. However, if the rate ratios between successive categories are similar, the differences between successive log(rates) should be constant. Therefore if you plot the rates on a log scale you will see a linear relationship with age. Type:

```
strate agecat, per(1000) graph yscale(log) ylab(50(50)150, angle(horizontal)) xlab(0/5, value(label))
```
The graph options used are as follows:

- `yscale(log)` specifies that the y-axis should be on the log scale.
- `ylab(50(50)150, angle(horizontal))` specifies y-axis labels, between 50 and 150 in units of 50, angle of text set to horizontal (default is vertical).
- `xlab(0/5, valuelabel)` specifies that label names should be displayed for age categories on the x-axis.

You can see that the trend in rates across age is approximately linear on the log scale. This indicates that the rate ratio from one age group to the next is similar.

### Comparison between groups: rate ratios

To obtain rate ratios for the effect of age at entry with the youngest age group as the baseline use the `stmh` command. You need to use the option `c(…)` to compare each of the higher categories to the baseline group.

```
stmh agecat, c(1,0)
```

Maximum likelihood estimate of the rate ratio comparing `agecat==1` vs. `agecat==0`

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.174</td>
<td>0.40</td>
<td>0.5250</td>
<td>0.715 1.927</td>
</tr>
</tbody>
</table>

```
stmh agecat, c(2,0)
```

© London School of Hygiene & Tropical Medicine 2011
Maximum likelihood estimate of the rate ratio comparing agecat==2 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.773</td>
<td>21.05</td>
<td>0.0000</td>
<td>1.760 4.371</td>
</tr>
</tbody>
</table>

stmh agecat, c(3,0)

Maximum likelihood estimate of the rate ratio comparing agecat==3 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.578</td>
<td>55.78</td>
<td>0.0000</td>
<td>2.952 7.101</td>
</tr>
</tbody>
</table>

stmh agecat, c(4,0)

Maximum likelihood estimate of the rate ratio comparing agecat==4 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.688</td>
<td>93.81</td>
<td>0.0000</td>
<td>4.286 10.438</td>
</tr>
</tbody>
</table>

stmh agecat, c(5,0)

Maximum likelihood estimate of the rate ratio comparing agecat==5 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.111</td>
<td>210.78</td>
<td>0.0000</td>
<td>10.114 28.949</td>
</tr>
</tbody>
</table>

It is clear that the rate ratios for consecutive age-groups versus the youngest age-group (baseline) increase with age.

The rate for the 45-49-year age-group is 1.17 times that of the 40-44 year age-group;
The rate for the 50-54-year age-group is 2.77 times that of the 40-44 year age-group;
The rate for the 55-59-year age-group is 4.58 times that of the 40-44 year age-group;
and so on.

To compare the rate ratios for consecutive age-groups you can examine the 95% confidence intervals. If they do not overlap then you can say that the two rate ratios are significantly different. However, if you examine the 95% confidence intervals from one age-group to the next, you can see that they overlap.
Stratified estimates for the exposure of interest

Let’s say the main exposure of interest is grade of employment (coded: 1 = high grade, 2 = low grade). You will first examine the all-cause mortality rates for low and high grades of employment, then use the `stmh` command to estimate the rate ratio for low-grade employees versus high-grade employees.

(strate grade, per(1000))

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)

<table>
<thead>
<tr>
<th>grade</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>221</td>
<td>20.3398</td>
<td>10.8654</td>
<td>9.5233</td>
<td>12.3966</td>
</tr>
<tr>
<td>2</td>
<td>182</td>
<td>7.2656</td>
<td>25.0496</td>
<td>21.6624</td>
<td>28.9665</td>
</tr>
</tbody>
</table>

(stmh grade, c(2,1))

Maximum likelihood estimate of the rate ratio comparing grade==2 vs. grade==1
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.305</td>
<td>73.78</td>
<td>0.0000</td>
<td>1.895 2.805</td>
</tr>
</tbody>
</table>

To assess for potential confounding or effect modification by age at entry on the effect of employment grade use `stmh` with `grade`, stratified by `agecat`. Type:

(stmh grade, by(agecat) c(2,1))

Maximum likelihood estimate of the rate ratio comparing grade==2 vs. grade==1 by agecat
RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>agecat</th>
<th>RR</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-</td>
<td>1.22</td>
<td>0.42</td>
<td>3.57</td>
</tr>
<tr>
<td>45-</td>
<td>1.36</td>
<td>0.67</td>
<td>2.75</td>
</tr>
<tr>
<td>50-</td>
<td>1.92</td>
<td>1.23</td>
<td>3.01</td>
</tr>
<tr>
<td>55-</td>
<td>1.43</td>
<td>1.00</td>
<td>2.06</td>
</tr>
<tr>
<td>60-</td>
<td>1.21</td>
<td>0.82</td>
<td>1.80</td>
</tr>
<tr>
<td>65-</td>
<td>1.40</td>
<td>0.54</td>
<td>3.62</td>
</tr>
</tbody>
</table>

© London School of Hygiene & Tropical Medicine 2011
To examine whether the crude RR estimate for low grade, RR=2.31 (95% CI 1.90 to 2.81), is confounded by age at entry, examine the age-specific estimates. First you should assess whether there is any effect modification (also called ‘interaction’). The test for effect modification is not significant (P=0.79) and the values of the age-specific RRs are quite similar (they range from 1.2 to 1.9 with no evidence of a trend). Therefore you can summarise them using the Mantel-Haenszel estimate, RR=1.43 (95% CI 1.16 to 1.76). However, the results do show that the crude estimate of the effect of grade was strongly confounded by age at entry (RR\textsubscript{crude} =2.31 compared to RR\textsubscript{MH}=1.43).

**Key points**

1. To analyse data from cohort studies using Stata, you first need to define the variables that identify the time scale and the outcome of interest with the command `stset`.

2. The effect of a risk factor is usually assessed in terms of rate ratios (RRs). We use the command `stmh` to do this. Crude estimates of a RR, however, should always be interpreted with caution, since other factors, related to the risk factor and associated with the outcome, may contribute to its value.

3. To assess for the potential confounding of an estimated RR you should:
   - stratify the data by categories of the confounder and compare the stratum-specific estimates of the RR for the risk factor.
   - if the stratum-specific estimates of the RR for the risk factor do not substantially differ from each other you can use the Mantel-Haenszel summary estimate to report the effect adjusted for the potential confounder; if the adjusted RR differs considerably from the crude RR we say that there is evidence of confounding.
   - If the stratum-specific estimates of the RR for the risk factor differ from each other you cannot use the Mantel-Haenszel summary estimate and the stratum-specific RRs should be reported.

**Review exercise**

Now try to carry out a similar analysis on your own. The solutions to the questions below are given in Section 4 of this workbook. In order to work through this exercise you need to work on the original dataset `whitehal.dta`.

1. In order to analyse CHD mortality, set the time and the CHD mortality outcome variables with `stset` (remember that the time variables are `timein` and `timeout`, the outcome is `chd`, the identifier is `id`, and the scale should be set to years). Use the `tab` command on the `chd` variable to check that the number of deaths from CHD corresponds with the number given in the `stset` output.
2 Recode age at entry into 5-year age-groups and investigate how CHD mortality varies according to age at entry to the study using `strate`.

3 Examine the CHD mortality rates for low and high grades of employment. Use the `stmh` command to estimate the rate ratio for low-grade employees versus high-grade employees.

4 Use `stmh` to examine the effect of `grade`, stratified by age at entry:
   Is there any evidence of interaction between employment grade and age at entry?
   Examine the result of the test for interaction.
   Is the effect of `grade` confounded by age at entry?
Section 3: Practical sessions

Practical 3

Survival analysis

Aim

To learn how to compute lifetables and Kaplan-Meier survival curves and test for a difference in the survival probabilities for different groups.

Objectives

By the end of this session you will be able to:

- produce lifetables and corresponding survival curves
- estimate a Kaplan-Meier survival curve
- compare survival curves using the Logrank test.

Planning your session

To complete the work in this session:

- You should have worked through Session 3 of *Statistical Methods in Epidemiology*
- The specific dataset for this session is `trinidad.dta`
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

```
e.g. cd c:\sme
```

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

```
e.g. log using smeprac3.log
```

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the `cd` command.
Remember:

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.

Note:

All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0 or older, and are compatible with later versions of Stata.

Reading in the dataset and identifying relevant variables

Throughout this session you will be analysing a dataset from a cohort study of 318 men carried out in Trinidad.

To read in the dataset, type:

use trinidad

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

help trinidad

You will look at survival within this cohort of men with respect to whether or not they had heart disease at the start of the study and smoking habit. The outcome of interest is death from any cause.

dath is the variable name for death from any cause coded: 0=no,1=yes
chdstart is the variable name for heart disease at time of entry coded: 0=no,1=yes
current the variable name for current smoking status coded: 0=current non-smoker, 1=current smoker
years is the variable containing years of follow-up

To find out how many deaths from any cause occurred during the follow-up period, type:

tab death

<table>
<thead>
<tr>
<th>Died from any cause</th>
<th>Preq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>230</td>
<td>72.33</td>
<td>72.33</td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td>27.67</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>318</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

There were a total of 88 deaths (28%) from any cause.
To find out how many men entered the study with CHD type:

**tab chdstart**

<table>
<thead>
<tr>
<th>Heart disease at</th>
<th>Freq.</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>time of entry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>252</td>
</tr>
<tr>
<td>86.90</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>100.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>290</td>
</tr>
</tbody>
</table>

To analyse data from cohort studies focussing on the time to an event you can use either the lifetable or the Kaplan-Meier method. Both methods can give a graphical description of the survival pattern of the cohort.

The lifetable method uses data summarised by intervals of time, e.g. yearly, whereas the Kaplan-Meier method uses the individual data when an event or censoring occurs.

**Lifetable method**

To produce the lifetable estimates by years of follow-up for men who entered the study with heart disease, you can use the `ltable` command. You need to specify the time of follow-up, the outcome variable and the criteria for the group of interest. Type:

`ltable years death if chdstart==1`

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg. Total</th>
<th>Deaths</th>
<th>Lost</th>
<th>Survival</th>
<th>Std. Error</th>
<th>[95% Conf. Int.]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>38</td>
<td>3</td>
<td>0</td>
<td>0.9211</td>
<td>0.0437</td>
<td>0.7749</td>
</tr>
<tr>
<td>1</td>
<td>35</td>
<td>2</td>
<td>0</td>
<td>0.8684</td>
<td>0.0548</td>
<td>0.7123</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>2</td>
<td>0</td>
<td>0.8158</td>
<td>0.0629</td>
<td>0.6521</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>3</td>
<td>0</td>
<td>0.7368</td>
<td>0.0714</td>
<td>0.5661</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>2</td>
<td>0</td>
<td>0.6842</td>
<td>0.0754</td>
<td>0.5115</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>0</td>
<td>3</td>
<td>0.6842</td>
<td>0.0754</td>
<td>0.5115</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>2</td>
<td>3</td>
<td>0.6206</td>
<td>0.0807</td>
<td>0.4430</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>0</td>
<td>6</td>
<td>0.6206</td>
<td>0.0807</td>
<td>0.4430</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>1</td>
<td>6</td>
<td>0.5516</td>
<td>0.0968</td>
<td>0.3462</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.5516</td>
<td>0.0968</td>
<td>0.3462</td>
</tr>
</tbody>
</table>

Lifetable output:

- The first two columns give the interval of follow-up in which the survival probabilities are calculated. So in the first year of follow-up the interval is 0 to 1.
- The column ‘Beg. Total’ gives the total number of men in the study at the beginning tart of each time interval.
- The ‘Deaths’ column shows the number of men who died during the interval.
- The ‘Lost’ column give the number of men lost to follow-up during the interval.
- The column ‘Survival’ gives the proportion surviving at the end of each interval, whereas the last 3 columns give the corresponding standard error and 95% confidence intervals.
So in the 7th year, i.e. the interval from 6 to 7:

There were 23 men in the study out of the 38 men who entered the study with heart disease remaining alive. During this year, 2 men died and 3 were lost to follow-up. The probability of survival for this group at the end of the year is calculated as follows:

\[
\text{Prob (death)} = \frac{\text{Number of deaths}}{\text{number at risk} - (0.5 \times \text{number censored})}
\]

\[
= \frac{2}{23 - (0.5 \times 3)}
\]

\[
= 0.093
\]

\[
\text{Prob (survival in year 7)} = 1 - \text{Prob (death)} = 0.907
\]

Finally, this probability is multiplied by the probability of surviving up to to the end of year 6.\((=0.6842)\)

So, the probability of survival for this group at the end of the year 7 is given by \(0.907 \times 0.6842 = 0.6206\).

If you add the option graph to ltable command you will obtain the corresponding survival plot. Type:

\[
ltable \text{ years death if chdstart==1, graph xlab(0(2)10) ylab(0.5(0.1)0.9, angle(horizontal)) noconf}
\]

The graph options used are as follows:

- \text{xlab(0(2)10)} specifies x-axis labels, between 0 and 10 years in 2-year units
- \text{ylab(0.5(0.1)0.9, angle(horizontal))} specifies y-axis labels, between 0.5 and 0.9 in units of 0.1, angle of text set to horizontal (default is vertical)
- \text{noconf} suppresses confidence intervals for survival probabilities, which are shown by default

From the plot you can see that by the end of the 3rd year the proportion surviving was over 80%. By the end of follow-up at 10 years the proportion surviving was 55%.
To compare the survival pattern of men who entered the study with heart disease and men who did not, you can produce a stratified lifetable. Type:

**ltable years death, by(chdstart)**

and to obtain the corresponding survival curves add the `graph` option. Type:

**ltable years death, by(chdstart) graph ylab(0.5(0.1)1.0, angle(horizontal)) noconf overlay**

The survival for men with heart disease at the start of the study was always lower than men without heart disease at the start of the study.
**Kaplan-Meier survival probabilities**

To obtain Kaplan-Meier estimates of these survival probabilities you must first define the follow-up information using the `stset` command. You must be careful to specify when individuals first became at risk using the `origin` option and specify the start of follow-up using the `enter` option. Type:

```
stset timeout, fail(death) enter(timein) origin(timein) id(id) scale(365.25)
```

```
    id:  id
    failure event:  death != 0 & death < .
    obs. time interval:  (timeout[_n-1], timeout]
    enter on or after:  time timein
    exit on or before:  failure
    t for analysis:  (time-origin)/365.25
    origin:  time timein
```

```
318  total obs.
    0  exclusions
```

```
318  obs. remaining, representing
318  subjects
88  failures in single failure-per-subject data
2204.539  total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 9.798768
```

To produce a Kaplan-Meier plot for the survival of men with heart disease at entry to the study, type:

```
sts graph if chdstart==1, ylab(0(0.25)1.0, angle(horizontal))
```

![Kaplan-Meier survival estimate](image)

Notice this plot is a step function, unlike the lifetable curve, which is smooth because it is calculated for specific intervals of time. Approximately, what is the survival probability at 10 years for men who enter the study with heart disease?

From the plot we can see this is about 0.55.
To produce the Kaplan-Meier survival curves for men with and without heart disease at entry to the study you must specify `chdstart` as the stratifying variable. Type:

```
sts graph, by(chdstart) ylab(0(0.25)1.0, angle(horizontal))
```

The two survival curves are presented on the same plot and show that the men who entered the study without heart disease have higher cumulative survival probabilities. To test this difference statistically you can use the Logrank test.

**Logrank test**

To test for a significant difference between two survival curves you simply use the command `test` with the `sts` command and specify the variable with the groups you want to compare. So, to compare the survival probabilities of men with and without heart disease at entry to the study, type:

```
sts test chdstart
```

<table>
<thead>
<tr>
<th>chdstart</th>
<th>Events</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>observed</td>
<td>expected</td>
</tr>
<tr>
<td>0</td>
<td>59</td>
<td>65.92</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>8.08</td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td>74.00</td>
</tr>
</tbody>
</table>

The result of the Logrank test shows that the two groups of men have significantly different cumulative survivals, \( P=0.01 \). However, you cannot quantify the difference in survival probabilities.

Remember when you have finished the session to close your log file, type:

```
log close
```
Review exercise

Now try to carry out a similar analysis on your own. The solutions to the questions below are given in Section 4. In order to work through this exercise you should work on the dataset trinidad.dta.

For this exercise:
- the outcome of interest is deaths from any cause
- the exposure of interest is current smoking habit

1. Obtain a lifetable for current smokers only. How many deaths from any cause were there in the 5th year? What was the probability of survival at the end of this year?
2. Compare the survival pattern of current smokers to that of current non-smokers in a stratified lifetable. What is the cumulative survival for both groups at the end of follow-up?
3. Obtain Kaplan-Meier survival curves for current smokers and current non-smokers. Is the survival probability for current smokers always less than that for current non-smokers?
4. Use a Logrank test to test the significance of the difference in the survival curves you produced in 3. Interpret your output.
Practical 4

Case-control studies

**Aim**

To learn how to analyse case-control data, obtain crude and adjusted estimates and test for trend with increasing exposures.

**Objectives**

By the end of this session you will be able to:

- use the commands `tab` and `mhodds` to examine the association between an exposure variable and case/control status in a case-control study, in both a crude analysis and stratifying by a potential confounder.
- use the `mhodds` command to investigate whether the effect of the main exposure of interest depends on value the of a second variable (interaction).

**Planning your session**

To complete the work in this session:

- you should have worked through Session 4 of *Statistical Methods in Epidemiology*
- The specific dataset for this session is `mwanza.dta`
- you must be familiar with using Stata and know the basic commands.

**Getting started**

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

```
  e.g. cd c:sme
```

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

```
  e.g. log using smeprac4.log
```

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the `cd` command.
**Remember:**

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.

**Note:**

All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0, and are compatible with later versions of Stata.

---

**Reading in the dataset and identifying relevant variables**

This practical session uses the dataset from Mwanza, Tanzania on HIV infection among women.

To read in the dataset, type:

```
use mwanza
```

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

```
help mwanza
```

You will look at the association between HIV infection and exposure to formal education, number of sexual partners and religion.

- **case** is the variable name for HIV infection coded: 1=case; 0=control
- **age1** is a grouped age variable coded: 1=15-19, 2=20-24, 3=25-29, 4=30-34, 5=35-44, 6=45-54 years
- **ed** is the variable name for level of education coded: 1=no formal education (none/adult only), 2=1-3 years, 3=4-6 years, 4=7+ years
- **npa** is the variable name for number of sexual partners ever coded: 1=0-1, 2=2-4, 3=5-9, 4=10-19, 5=20-49, 6=50+, 9=missing
- **rel** is the variable name for type of religion coded: 1=Moslem, 2=Catholic, 3=Protestant, 4=other, 9=missing

To examine how many cases and controls there are in the dataset, type:

```
tab case
```
To look at exposure to formal education create a new variable `ed2` which takes the value 1 for women with no formal education and value 2 for those with some education. Type:

```
gen ed2 = ed
recode ed2 3/4=2
```

To check that the new variable has been coded correctly, tabulate it against the original variable. Type:

```
tab ed ed2
```

```
<table>
<thead>
<tr>
<th>ed2</th>
</tr>
</thead>
</table>
| Education |         1          2 |     Total
|-----------|
| 1 |       312          0 |       312
| 2 |         0        75 |        75
| 3 |         0        365 |       36
| 4 |         0         11 |        11
|-----------|
| Total |       312        451 |       763
```

Similarly for age, recode `age1` to a new variable `age2` with the 4 categories: 1=15-19, 2=20-29, 3=30-44, 4 = 45+ years. Type:

```
gen age2 = age1
recode age2 3=2 4/5=3 6=4
```

Again you should tabulate the old variable against the new variable to check the coding is correct.

```
tab age1 age2
```

```
<table>
<thead>
<tr>
<th>age2</th>
</tr>
</thead>
</table>
| Age group |         1          2          3          4 |     Total
|-----------|
| 1 |       109          0          0          0 |       109
| 2 |         0        165          0          0 |       165
| 3 |         0        123          0          0 |       123
| 4 |         0          0       118          0 |       118
| 5 |         0          0       137          0 |       137
| 6 |         0          0          0       111 |       111
|-----------|
| Total |       109        288        255        111 |       763
```
Crude odds ratio estimate

To examine the relationship between being a case and formal education, type:

```
tab case ed2, row chi
```

<table>
<thead>
<tr>
<th>Case/control</th>
<th>ed2</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>263</td>
<td>311</td>
<td>574</td>
</tr>
<tr>
<td></td>
<td>45.82</td>
<td>54.18</td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td>49</td>
<td>140</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>25.93</td>
<td>74.07</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>312</td>
<td>451</td>
<td>763</td>
</tr>
<tr>
<td></td>
<td>40.89</td>
<td>59.11</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Pearson chi2(1) = 23.2789 Pr = 0.000

Note: You should examine the row percentages because column percentages are affected by the different probabilities of selection for cases and controls.

You must be clear about which variable you are treating as the exposure and which category is a case in your interpretation of the table. Examine the table above. What is the proportion of cases with some formal education?

There are 140/189 cases with some formal education, i.e. 74.1%.

To produce an odds ratio for exposure to formal education you can use the `mhodds` command. Try the following command first:

```
mhodds case ed2, c(1,2)
```

Maximum likelihood estimate of the odds ratio
Comparing ed2==1 vs ed2==2

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.413878</td>
<td>23.25</td>
<td>0.0000</td>
<td>0.285782 0.599391</td>
</tr>
</tbody>
</table>

Now change the baseline for `ed2`. Type:

```
mhodds case ed2, c(2,1)
```

Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs ed2==1

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.416169</td>
<td>23.25</td>
<td>0.0000</td>
<td>1.668360 3.499168</td>
</tr>
</tbody>
</table>

The P-value shows strong evidence against the null hypothesis of no association.

The first version of the command takes level 2 (some education) as baseline. Thus, the odds ratio is 1.0 divided by the odds ratio from the second version, which uses level 1 (no formal education) as the baseline. It is important to know which level is the baseline in your interpretation of the odds ratio.
Adjusted odds ratio estimates

Now you will examine the effect of education on HIV infection adjusted for age. To produce tables of case status by education stratified by age, you can use the `by()` option before the `tab` command, but first you must sort the data by age2. Type:

```
sort age2
```

```
by age2: tab case ed2
```

Note: These tables have cases with the exposure in the bottom right corner, not in the top left corner.

To obtain the odds ratio for HIV infection, comparing those with and without education within each stratum, use `mhodds` with the option `by(age2)`. Do you think it’s appropriate to produce a summary estimate of the odds ratio adjusted for age?
mhodds case ed2, by(age2) c(2,1)

Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs ed2==1
by age2

<table>
<thead>
<tr>
<th>age2</th>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.519231</td>
<td>1.02</td>
<td>0.3132</td>
<td>0.142000 1.898596</td>
</tr>
<tr>
<td>2</td>
<td>1.952381</td>
<td>4.10</td>
<td>0.0430</td>
<td>1.009050 3.777605</td>
</tr>
<tr>
<td>3</td>
<td>3.458647</td>
<td>16.76</td>
<td>0.0000</td>
<td>1.836539 6.513465</td>
</tr>
<tr>
<td>4</td>
<td>2.847222</td>
<td>3.05</td>
<td>0.0808</td>
<td>0.833046 9.731360</td>
</tr>
</tbody>
</table>

Mantel-Haenszel estimate controlling for age2

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.289748</td>
<td>17.94</td>
<td>0.0000</td>
<td>1.543506 3.396777</td>
</tr>
</tbody>
</table>

Test of homogeneity of ORs (approx): chi2(3) = 8.03
Pr>chi2 = 0.0455

The χ² value for effect modification suggests that there is a different effect of education on HIV infection depending on age. The confidence intervals of the ORs are wide but there is an indication that education may be protective in the youngest age group, or at least not as "harmful".

If you judge that the interaction is real, the combined estimate of 2.29 should not be used and stratum-specific estimates presented. It is plausible that the "effect" of education has changed if there has been awareness of HIV risks and teaching about this in schools in recent years.

Test for trend

To look for evidence of a dose-response effect of years of schooling on HIV infection, you can use tabodds to perform a test for trend. Type:

```
tabodds case ed
```

<table>
<thead>
<tr>
<th>ed</th>
<th>cases</th>
<th>controls</th>
<th>odds</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49</td>
<td>263</td>
<td>0.18631</td>
<td>0.13734 0.25275</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>51</td>
<td>0.47059</td>
<td>0.28969 0.76444</td>
</tr>
<tr>
<td>3</td>
<td>110</td>
<td>255</td>
<td>0.43137</td>
<td>0.34495 0.53945</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>5</td>
<td>1.20000</td>
<td>0.36623 3.93196</td>
</tr>
</tbody>
</table>

Test of homogeneity (equal odds): chi2(3) = 26.70
Pr>chi2 = 0.0000

Score test for trend of odds: chi2(1) = 22.24
Pr>chi2 = 0.0000

As a result of the first test (test for homogeneity) you can reject the null hypothesis that the odds of HIV infection are the same in each education category. As a result of the second test (score test for trend) you can say that there is some evidence for a trend of increasing odds of HIV infection with increasing years of education.
Remember that this is a case-control study, so that D/H or case/control does not give you the exact odds because the probabilities of selection differ between cases and controls. However the “odds” column is a constant multiple of the true odds so can be used to look at trends.

Repeat the same analysis using `mhodds`, type:

```
mhodds case ed
```

Score test for trend of odds with ed
(The OR estimate is an approximation to the odds ratio for a one unit increase in ed)

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.504866</td>
<td>22.24</td>
<td>0.0000</td>
<td>1.269793 1.783457</td>
</tr>
</tbody>
</table>

Note: The $\chi^2$ value is the same as for the score test above. The estimate of the odds ratio tells us that for each increase in education category the odds of having HIV infection is multiplied by 1.5. To be more specific, in Mwanza the odds of having HIV infection for women with 1-3 years of education is 1.5 times that of women with no education; the odds of having HIV infection for those who have 4-6 years of education is 1.5 times that of those with 1-3 years of education.

To investigate further whether there really is evidence that risk of HIV infection increases with years of schooling, perform a test for trend excluding women who had never been to school. To exclude women with no formal education, type:

```
mhodds case ed if ed>1
```

Score test for trend of odds with ed
(The OR estimate is an approximation to the odds ratio for a one unit increase in ed)

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.119786</td>
<td>0.21</td>
<td>0.6458</td>
<td>0.691185 1.814162</td>
</tr>
</tbody>
</table>

What do you conclude about the association between schooling and HIV infection?

There is no evidence that there is a significant trend in the odds ratios for women with some formal education.
Key points

- obtain a crude estimate first and then see how it compares with adjusted estimates to decide whether there is confounding
- know your data and how to deal with the unknown values
- it is a matter of judgement whether confounding or interaction is present even though there is a statistical test for the latter
- ensure that comparisons of crude and adjusted estimates cover exactly the same group of people
- tabodds can be used to look at trends but does not give exact odds in case-control studies.

Review exercise

Now try to carry out a similar analysis on your own. The solutions to the questions below are given in Section 4. In order to work through this exercise you should work on the dataset mwanza.dta.

1. Investigate whether religion (rel) confounds the association between schooling (ed2) and HIV infection (case). Note: rel has a code 9 for missing values, so we suggest you set this to system-missing (.).
   
   Hint: To confine the adjusted MHOR estimate to those cases with known religion, type:
   
   `mhodds case ed2 if rel!=., by(rel) c(2,1)`

2. You might expect an increasing risk of HIV infection with number of sexual partners. Carry out a test for trend using npa and estimate the odds ratio for each increase in category of number of partners.
Practical 7

Logistic regression 1

Aim

To learn how to carry out a simple logistic regression analysis.

Learning objectives

By the end of this session you will be able to:

- obtain and interpret a logistic model for a binary exposure variable
- obtain and interpret a logistic model for an exposure with more than 2 levels
- use a likelihood ratio test to assess the contribution of a variable to a model
- obtain a logistic model with the effect of one exposure, controlled for the effect of others
- use a likelihood ratio test to test the significance of the effect of one exposure, controlled for the effect of a second variable
- assess the confounding effect of variables within a model.

Planning your study

To complete the work in this session:

- You should have worked through Session 7 of Statistical Methods in Epidemiology
- The specific dataset for this session is oncho.dta
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

e.g. cd c:sme

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

e.g. log using smeprac7.log

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the cd command.
Remember:
1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.

Note:
All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0, and are compatible with later versions of Stata.

Reading in the dataset and identifying relevant variables

In this practical session you will use a dataset from the prevalence study of onchocerciasis in Sierra Leone.

To read in the dataset, type:
```
use oncho
```

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:
```
help oncho
```

In this analysis you will look at the association between microfilariae infection and exposure to living in the savannah region or the forest region. You will also consider how (if at all) the association changes with age and sex.

- **mf** is the variable name for microfilarial infection coded: 0=no, 1=yes
- **area** is the variable name for area of residence coded: 0=savanna, 1=forest
- **agegrp** is the variable name for age-group coded: 0=5-9, 1=10-19, 2=20-39, 3=40+
- **sex** coded: 0=male, 1=female

Use commands **describe**, **summarise** and **list** to examine the data. You can stop listing by holding down the control key and tapping the break key (ctrl+break).

To produce frequency distributions for **mf**, **area** and **agegrp** use **tab**, type:
```
tab mf
```
### Practical 7: Logistic regression 1

#### Testing for an association

For an initial examination of the association between infection and area of residence use the `tab` command. Type:

```
tab mf area, col chi
```

<table>
<thead>
<tr>
<th>Microfil. infection</th>
<th>Area of residence</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>267</td>
<td>213</td>
<td>480</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>281</td>
<td>541</td>
<td>822</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>548</td>
<td>754</td>
<td>1302</td>
<td></td>
</tr>
</tbody>
</table>

Pearson chi2(1) = 57.1512  Pr = 0.000

There were 822 individuals with microfilarial infection.

754 individuals lived in the forest area and 548 lived in the savannah area.

There were 202 children aged 5 to 9 years; 218 individuals aged 10 to 19 years; 424 individuals aged 20 to 39 years; and 458 individuals aged 40+ years.

From the table you can see that 51% of individuals living in the savannah were infected compared to 72% of individuals living in the forest. This is a strongly significant difference P<0.001.
To examine the odds of infection for each area of residence use `tabodds`, type:

```
tabodds mf area
```

<table>
<thead>
<tr>
<th>area</th>
<th>cases</th>
<th>controls</th>
<th>odds</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>281</td>
<td>267</td>
<td>1.05243</td>
<td>0.89012</td>
</tr>
<tr>
<td>1</td>
<td>541</td>
<td>213</td>
<td>2.53991</td>
<td>2.16752</td>
</tr>
</tbody>
</table>

Test of homogeneity (equal odds): chi2(1) = 57.11  Pr>chi2 = 0.0000

You can see that the odds of infection is significantly greater in the forest area compared to the savannah.

Note: the columns are headed ‘cases’ and ‘controls’ by default, even though the variable here is disease status.

Use `mhodds` to obtain an odds ratio estimate. Type:

```
mhodds mf area, c(1,0)
```

Maximum likelihood estimate of the odds ratio
Comparing area==1 vs area==0

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.413363</td>
<td>57.11</td>
<td>0.0000</td>
<td>1.906097</td>
</tr>
</tbody>
</table>

Therefore, the odds of infection in the forest area is 2.4 times that in the savannah area (P<0.001, 95% CI 1.91 to 3.06).

**Logistic regression with one binary exposure**

Now you will produce the same result but using logistic regression. To obtain a logistic model on a log scale you must use the `logit` command (as opposed to `logistic`, which presents the estimates on the original odds ratio scale). The `logit` command gives the parameter estimates for log odds.

The first model you will fit is:

```
log odds = constant + area
```

Type:

```
logit mf area
```
Note that the output above the table, logistic regression estimates are derived by starting with a guess of the parameter estimates, then using the result to compute a better guess (nearer to the maximum likelihood estimates). This is known as ‘iteration’. The log likelihood at each iteration is shown. The procedure stops when there is no further increase in the log likelihood. Note also that the log likelihood for the model is given above the table.

The statistics in the column above the table on the right are:
1 the number of observations.
2 a likelihood ratio $\chi^2$ test for the null hypothesis that none of the variables in the model are associated with the outcome variable. In this instance there is only one variable (area) in the model, so this is a test of the null hypothesis that area is not associated with microfilarial infection.
3 the P value for the likelihood ratio test.
4 a ‘goodness of fit’ statistic (we are not interested in this for the purposes of this course).

Now consider the values in the table. The first column gives the parameter names for each row of estimates in the table. The model corresponds to:

Log odds = constant + area

The values in the coefficient column represent the log (OR) for the effect of forest (0.881) and the constant log odds in the savannah area (0.051).

Log odds = 0.051 + 0.881 area

The third column gives the standard error for the model coefficients. These are then used to calculate z, the Wald test statistic, the corresponding P-value and finally the 95% confidence limits. Consider the estimates for area. What can you conclude from these values about the effect of forest on odds of infection?

log OR = 0.881  (95% CI: 0.65 to 1.11)
Wald test:  z = 7.49,  P<0.001
You can see there is a highly significant association between odds of infection and area of residence. You can be 95% confident that the true log (OR) does not equal zero (therefore the true OR does not equal 1).

To obtain the OR estimate for the effect of living in the forest area you should take the exponential of the coefficient = \( \exp(0.881) = 2.41 \).

The log scale is preferable for explaining the model estimates and how 95% confidence intervals and Wald tests are derived. However, Stata allows you to automatically obtain estimates on the odds ratio scale, which is convenient for reporting results.

Type:

logistic mf area

| mf       | Odds Ratio | Std. Err. | z    | P>|z|    | [95% Conf. Interval] |
|----------|------------|-----------|------|--------|----------------------|
| area     | 2.413363   | .2839914  | 7.487| 0.000  | 1.916275 3.039397    |

The results are now shown as odds ratios. Note that:

- The constant term (_cons) is omitted when the logistic command is used.
- The standard error for the odds ratio is only approximate and should be ignored.
- The z statistic (for the Wald test) is identical to that when the logit command is used. This is derived using the standard error for the log odds ratio.
- Similarly, the confidence intervals are also derived using the standard error for the log odds ratio.

You can also obtain the same results using the or option with the logit command (where or = odds ratio estimates). Type:

logit mf area, or
Exposures with more than 2 levels

Use the `tab` command to examine the association between age-group and microfilarial infection. Type:

```
tab mf agegrp, col
```

<table>
<thead>
<tr>
<th>Microfilarial infection</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>156</td>
<td>119</td>
<td>125</td>
<td>80</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td>77.23</td>
<td>54.59</td>
<td>29.48</td>
<td>17.47</td>
<td>36.87</td>
</tr>
<tr>
<td>1</td>
<td>46</td>
<td>99</td>
<td>299</td>
<td>378</td>
<td>822</td>
</tr>
<tr>
<td></td>
<td>22.77</td>
<td>45.41</td>
<td>70.52</td>
<td>82.53</td>
<td>63.13</td>
</tr>
<tr>
<td>Total</td>
<td>202</td>
<td>218</td>
<td>424</td>
<td>458</td>
<td>1302</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The prevalence of infection of cases increases from 23% in the 5-9 year age-group to 83% in the 40+ years age-group.

In order to produce an estimate for each level compared to the baseline, for exposures with more than 2 levels you need to tell Stata to create indicator variables. To do this you type `xi:` before the `logit` (or `logistic`) command and an `i.` before the variable with more than 2 levels. To produce a model for age-group, type:

```
xi: logit mf i.agegrp
```

```
i.agegrp     _Iagegrp_0-3     (naturally coded; _Iagegrp_0 omitted)
```

Iteration 0: log likelihood = -857.02925
Iteration 1: log likelihood = -729.02713
Iteration 2: log likelihood = -727.83463
Iteration 3: log likelihood = -727.83149

Logit estimates

```
Log likelihood = -727.83149                  Number of obs = 1302
LR chi2(3) = 258.40                         Prob > chi2 = 0.0000
Pseudo R2 = 0.1508
```

```
mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------
_Iagegrp_1    |   1.037211   .2159948     4.80   0.000     .6138689    1.460553
_Iagegrp_2    |   2.093344   .1987306    10.53   0.000      1.70384    2.482849
_Iagegrp_3    |   2.774082   .2080735    13.33   0.000     2.366266    3.181899
_cons         |  -1.221215   .1677778    -7.28   0.000    -1.550053    -0.892376
```

Before you look at this more closely, produce the output on the odds ratio scale.
Type:

```
 xi: logistic mf i.agegrp
```

|   | Odds Ratio   | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---|--------------|-----------|------|------|----------------------|
| _Iagegrp_1 | 2.821337 | .6093942 | 4.80 | .000 | 1.847566 - 4.308341 |
| _Iagegrp_2 | 8.112 | 1.612103 | 10.53 | .000 | 5.495005 - 11.97534 |
| _Iagegrp_3 | 16.02391 | 3.334152 | 13.33 | .000 | 10.65752 - 24.09246 |

Note that there are three odds ratios each of which refers to the same baseline group (those aged 5-9 years):

- the odds ratio is 2.82 for those aged 10-19 compared to those aged 5-9 years
- the odds ratio is 8.11 for those aged 20-39 compared to those aged 5-9 years
- and the odds ratio is 16.02 for those aged >40 compared to those aged 5-9 years.

There are three Wald test P-values (one for each odds ratio) which test whether each odds ratio is significantly different from 1. In this example, all three odds ratios are significant (P<0.001).

The likelihood ratio statistic is 258.4 on 3 degrees of freedom (P<0.001). Note that there is only one likelihood ratio test P-value. This tests the significance of the variable `agegrp`, by simultaneously testing the significance of the three parameters in the model (the estimates of the log odds ratios for age groups 1, 2 and 3 versus age group 0).

Note: Because area is a binary variable coded as 1 and 0, it makes no difference if it is used as a categorical variable. You can check this by typing the two commands below and comparing them:

```
xl: logistic mf i.area
logistic mf area
```

### Likelihood ratio test

This section shows how to carry out a likelihood ratio test in Stata. Remember, a likelihood ratio test compares a model with parameter(s) of interest to a model without parameter(s) of interest, to assess the contribution of a variable (or parameter) to the model.

As noted in CAL session 7, the likelihood ratio statistic (LRS) is calculated by using the difference between $L_1$ (the log likelihood when the exposure variable is included in the model) and $L_0$ (the log likelihood when the variable is excluded from the model):

$$LRS= 2(L_1 - L_0)$$

We then refer the LRS to the $\chi^2$ distribution on $(r-1)(c-1)$ degrees of freedom.
Note: the degrees of freedom is equal to the number of parameters excluded from the model when the exposure variable is excluded.

To compare the log likelihood from a model with agegrp \((L_1)\) and the log likelihood from the model without agegrp \((L_0)\), there are 5 steps involved:

1. fit the first model
2. save \(L_1\)
3. fit the second model
4. save \(L_0\)
5. compare \(L_0\) to \(L_1\)

To do this in Stata 8.0, type the following five commands:

\[\text{xi: logistic mf i.agegrp}\]

fit the model with agegrp

\[\text{estimates store A}\]

save \(L_1\) from this model, which is given in the model as Log likelihood = -727.83

\[\text{logistic mf}\]

fit the null hypothesis model, i.e. omit agegrp

\[\text{estimates store B}\]

save \(L_0\) from this model, which is given in the model as Log likelihood = -857.03

\[\text{lrtest A B}\]

compare \(L_0\) from this model with \(L_1\), i.e. \(LRS = 2(-727.83 \text{ minus } -857.03) = 258.40\)

Note: the command estimates can be shortened to est.

You should see the following output:

\[
\text{likelihood-ratio test} \\
\text{(Assumption: B nested in A)}
\]

\[
\begin{array}{ll}
\text{LR chi2(3) } & 258.40 \\
\text{Prob > chi2} & 0.0000
\end{array}
\]

This is the likelihood ratio test to assess the significance of agegrp in the model. As you can see it is highly significant, \(P<0.0001\).

Note: The two models being compared in the LRT must be fitted on exactly the same data. This may not happen if some observations have missing values for the variable being tested. Stata 8.0 will warn you if this is the case. To overcome this, you should fit all your models excluding missing values (i.e. the third command in the example above should be changed to: \text{logistic mf, if agegrp!=.}.

---

**Logistic regression with more than one exposure**

You can now look at how you can fit a model with two exposures. You will produce a logistic regression to obtain the odds ratios, confidence intervals and likelihood ratio statistics for:

- the effect of area controlling for agegrp, and
- the effect of agegrp controlling for area on the odds of microfilarial infection.
To display the odds ratio estimates, use the **logistic** command. Type:

```
xi: logistic mf area i.agegrp
```

<table>
<thead>
<tr>
<th>i.agegrp</th>
<th>_Iagegrp_0-3 (naturally coded; _Iagegrp_0 omitted)</th>
</tr>
</thead>
</table>

Logistic regression

- Number of obs = 1302
- LR chi2(4) = 329.24
- Prob > chi2 = 0.0000
- Log likelihood = -692.40733
- Pseudo R2 = 0.1921

| mf | Odds Ratio   | Std. Err. | z    | P>|z|     | [95% Conf. Interval] |
|----|--------------|-----------|------|---------|----------------------|
| area | 3.083224  | .424372   | 8.18 | 0.000   | 2.354217   | 4.037975            |
| _Iagegrp_1 | 2.599132 | .5771594  | 4.30 | 0.000   | 1.681945   | 4.016473            |
| _Iagegrp_2 | 9.76541   | 2.033437  | 10.94| 0.000   | 6.49301    | 14.68706            |
| _Iagegrp_3 | 17.64158  | 3.808709  | 13.29| 0.000   | 11.55496   | 26.93437            |

Have the coefficients changed from those in the models with each variable alone?

In the model with only area, the OR estimate for the effect of forest was 2.41. After controlling for age the OR increased to 3.08. Hence, age slightly confounded the effect of area. The OR estimates for age without adjusting for area were, 2.82, 8.11 and 16.02, so again there is only a slight difference in the estimates after controlling for age.

In the crude (unadjusted) analysis, living in the forest and older age are both risk factors for microfilarial infection. The estimated association between area and microfilarial infection increases when controlled for age group. Similarly, the estimated effect of the two older age group increases when controlled for area.

You should now use the likelihood ratio test

1. To test the significance of **area**, adjusted for **agegrp**
2. To test the significance of **agegrp** adjusted for **area**.

For the likelihood ratio test for **area**, type:

```
xi: logistic mf area i.agegrp
est store A
xi: logistic mf i.agegrp
est store B
lrtest A B
```

<table>
<thead>
<tr>
<th>likelihood-ratio test (Assumption: B nested in A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR chi2(1) = 70.85</td>
</tr>
<tr>
<td>Prob &gt; chi2 = 0.0000</td>
</tr>
</tbody>
</table>

Therefore, after adjusting for **agegrp**, **area** is still highly significant in the model.
For the likelihood ratio test for `agegrp`, type:

```
xii: logistic mf area i.agegrp
est store A
xii: logistic mf area
est store B
lrtest A B
```

likelihood-ratio test  LR chi2(3) = 272.22
(Assumption: B nested in A)  Prob > chi2 = 0.0000

Therefore, after adjusting for `area`, `agegrp` is still also highly significant in the model.

---

**Key points**

1. Before doing logistic regression, we should produce tables of explanatory variables by disease with appropriate percentages. For example, a table of `mf` by `agegrp` showed us that the prevalence of `mf` increased with age from 23% in those aged 5-9 years to 83% in those aged ≥40 years.

2. Logistic regression models the log odds of disease and produces odds ratios. The outcome must be binary in order to use logistic regression to model it. You have used logistic regression to obtain an odds ratio (with 95% CI) for the binary variable `area`, and to obtain odds ratios (with 95% CI) for a variable with four levels `agegrp`.

3. The Wald test is based on the log OR divided by its SE. It tests whether an odds ratio is significantly different from 1. The Wald P-value may be obtained directly from the logistic regression output, but the SE is only approximate and should not be used.

4. The likelihood ratio test compares the log likelihood from the models with and without the variable of interest. It tests whether the variable of interest is significantly associated with the outcome. There is one Wald test for every odds ratio in the model and one LRT for every variable in the model, e.g. there was one LRT P-value for `agegrp` but three Wald P-values (one for each OR).

5. We have used logistic regression to obtain adjusted odds ratios. The odds ratios obtained from a logistic model are all adjusted for the other variables in the model. For example, a logistic model with `agegrp` and `area` as explanatory variables produces odds ratios for these two variables adjusted for each other.

6. We can use the likelihood ratio test to determine whether a variable has a significant effect on the outcome, after adjusting for the other variables in the model.

7. In order to determine whether a variable is a confounder, compare the crude (unadjusted) odds ratio with the adjusted odds ratio when these have been derived using the same method and the same observations.
Key points for Stata commands

1. Use the `logit` command to obtain the corner and log odds ratios; these may be used to obtain the odds of disease. Use the `logistic` command to obtain odds ratios. Most of the time, we will only be interested in odds ratios.

2. Indicator variables are used to obtain odds ratios for variables with more than two levels. Stata does this using `xi` and `i`, e.g. `xi: logistic mf i.agegrp`.

3. The two models being compared in the LRT must be fitted on exactly the same data. This may not happen if some observations have missing values for the variable being tested. To overcome this, fit all your models excluding missing values.

Review exercise

Now try to carry out the same analyses on your own. For this exercise you should use the `oncho` dataset. The solutions are given in Section 4.

1. Examine the association between microfilarial infection (`mf`) and `sex` using the `tab` and `logistic` commands.
   - Does it make any difference if you use `i.sex` instead of `sex` in the `logistic` command?
   - What is the OR estimate for `sex`?
   - Is there a significant association between microfilarial infection and `sex`?

2. Does the association between `sex` and microfilarial infection change when you control for the effects of `area` and `agegrp`?

3. Carry out a likelihood ratio test to assess whether `sex` should be included in the model with `area` and `agegrp`. 
Practical 8

Logistic regression 2

Aim

To learn how to examine and interpret interaction between two variables in a logistic model.

Learning objectives

By the end of this session you will be able to:

- investigate interaction between two exposure variables using logistic regression
- interpret the interaction parameters in your regression output
- perform a likelihood ratio test for interaction
- derive stratum-specific odds ratios when interaction is present.

Planning your study

To complete the work in this session:

- You should have worked through Session 8 of Statistical Methods in Epidemiology
- The specific dataset for this session is oncho.dta
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

e.g. cd c:sme

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

e.g. log using smeprac8.log

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the cd command.
Remember:

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.

2. Stata is case-sensitive and all commands should be typed in lower case.

3. In Stata output P=0.000 indicates a P-value of P<0.001.

Note:

All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0, and are compatible with later versions of Stata.

In this session the models you are instructed to produce are all on an odds ratio scale using the command `logistic`. In addition you can obtain the model estimates on a log scale using the `logit` command.

Reading in the dataset and identifying relevant variables

In this practical session you will again use the dataset from the prevalence study of onchocerciasis in Sierra Leone.

To read in the dataset, type:

```
use oncho
```

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

```
help oncho
```

In this session you will focus again on area of residence as the main exposure and microfilarial infection as the outcome. You should examine whether the effect of area on infection is modified by age. The aim of all statistical modelling exercises is to obtain the best measure of effect for the exposure of interest. There is no point in having a complex model which is difficult to interpret if a simpler one is statistically almost as good. So, your strategy should be to test a model with interaction between `area` and `agegrp`. If it does not significantly improve the model, tested via the likelihood ratio test (LRT), then a simpler model with no interaction should be chosen.

First a recap of the variables of interest:

- **mf** is the variable name for microfilarial infection coded: 0=no, 1=yes
- **area** is the variable name for area of residence coded: 0=savanna, 1=forest
- **agegrp** is the variable name for age-group coded: 0=5-9, 1=10-19, 2=20-39, 3=40+
- **sex** coded: 0=male, 1=female
To illustrate with a less complex model, use `recode` with age to create a binary variable. This is only less complex in that there will be fewer parameters in the model. (It is possible to test interactions with more than two categories in a variable, but we are trying here to keep things simpler at first.) To create a new variable where age takes 2 levels, type:

```
gen agebin = agegrp
recode agebin 0/1=0 2/3=1
```

Check the recoding of the new variable `agebin` by tabulating it against the original variable `agegrp`. Type:

```
tab agegrp agebin
```

```
agebin  
Age group |    0    |    1    | Total
-----------|--------|--------|-------
0  |  202   |   0    |  202
1  |  218   |   0    |  218
2  |    0   |  424   |  424
3  |    0   |  458   |  458
-----------|--------|--------|-------
Total    |  420   |  882   | 1302
```

### A logistic model with interaction

In Practical 7 you produced a logistic model that assumed proportional odds. This assumed that the effect of area on infection is the same for all levels of age. If the effect of area was modified by age, i.e. there was interaction between the two variables, you should include the interaction in your model.

To produce a model with interaction in Stata you put an asterisk (*) between the two variables for which the interaction parameter(s) are required. Type:

```
xi: logistic mf i.agebin*area
```

Note: When specifying interaction in a model like this you need to specify the indicator variables first, i.e. the variables with an `i.` in front.
Consider the output. Can you identify the following in the model?

- the main effect of type of area at the baseline value of age (lowest age group)
- the main effect of age group at the baseline value of area (savannah)
- the interaction terms between age and type of area.

1. The first row _agebin_1 refers to the main effect of agebin, that is the effect of age 20+ years compared to the baseline age of 5-19 years in the savannah (baseline for area).

2. The second row area refers to the main effect of area, that is the effect of forest compared to the savannah at the baseline of agebin, i.e. 5-19 year age-group.

3. The third row represents the interaction parameter. This is the additional effect for either variable when the other is not the baseline.

Therefore, the odds ratio for microfilarial infection in the forest area compared with savannah in the 5-19 year age-group is 2.42.

However, for exposure to the forest and age-group 20+ years compared with living in the savannah and being younger than 20 years, you must multiply the main effects and interaction estimates to obtain the appropriate odds ratio = 5.8 x 2.41 x 1.59 = 22.22.

**Likelihood ratio test for interaction**

Use a likelihood ratio test to assess the significance of the interaction. Type the following commands:

```
xi: logistic mf i.agebin*area
est store A
xi: logistic mf i.agebin area
est store B
lrtest A B
```

```
likelihood-ratio test               LR chi2(1) =  2.70
(Assumption: B nested in A)          Prob > chi2 = 0.1004
```

Do you think a model with interaction is better than one without interaction? The test for interaction is P=0.10, so we cannot reject the null hypothesis that the two models fit the data equally well. In other words, the interaction term does not contribute significantly to the fit of the model, and we can assume no interaction and the proportional odds model seems appropriate.
\textbf{Stratum–specific estimates}

If there is significant interaction, stratum-specific estimates for the effects of exposure on outcome should be presented. These can be ‘directly’ obtained in the following ways:

1 Stratum-specific odds ratios can be derived by multiplying the appropriate main effect and interaction terms in the standard interaction model, i.e. the model which includes a term for the interaction parameter(s).

For example, to derive the age-stratum-specific effect on infection of exposure to forest (compared with savannah) in the age-group 5-19 years, use only the main effect for area:

Forest, 5-19 years: $\text{OR} = 2.41$.

For the age-stratum-specific effect on infection of exposure to forest (compared with savannah) in the age-group 20+ years, you should combine the main effects of area and the interaction between agebin and area:

Forest, 20+ years: $\text{OR} = 2.41 \times 1.59 = 3.83$ (note effect of rounding coefficients before multiplying).

2 Stratum-specific estimates can also be obtained by producing a logistic model for each level of the stratifying variable. For example, to produce estimates for each level of agebin you would give the following commands:

\begin{verbatim}
logistic mf area if agebin==0
logistic mf area if agebin==1
\end{verbatim}
These commands would produce two models with estimates for each age-group.

**Age-group 5-19 years:**

**xi: logistic mf area if agebin==0**

| Odds Ratio | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|------------|-----------|-------|-------|---------------------|
| area       | 2.416252  | .5418676 | 3.934 | 0.000               | 1.556869, 3.750009 |

**Age-group 20+ years:**

**xi: logistic mf area if agebin==1**

| Odds Ratio | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|------------|-----------|-------|-------|---------------------|
| area       | 3.846787  | .6544816 | 7.919 | 0.000               | 2.755986, 5.369319 |

The disadvantage of this approach is that each model has a smaller sample size hence, less power, than a single model. The next two methods allow the full dataset to be used.

3 It is also possible to create new indicator variables using a method called ‘reparameterisation’ which directly give the stratum-specific estimates required:

```stata
gen areaage0 = agebin==0 & area==1
```

This will take the value 1 when area = 1 and agebin = 0, otherwise the value 0.

Similarly,

```stata
gen areaage1 = agebin==1 & area==1
```

This will take the value 1 when area=1 & agebin=1, otherwise 0.
To derive age-group-specific effects for area, fit the model:

**xi: logistic mf i.agebin areaage***

(areaage* is the same as areaage0 areaage1)

```
.i.agebin      _Iagebin_0-1 (naturally coded; _Iagebin_0 omitted)
Logistic regression                                Number of obs   =       1302
LR chi2(3)      =     300.08
Prob > chi2     =     0.0000
Log likelihood  = -706.99054                       Pseudo R2       =     0.1751
------------------------------------------------------------------
mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
   _Iagebin_1 |    5.80094   1.236674     8.25   0.000     3.819749    8.809715
  areaage0   |   2.416252   .5418676     3.93   0.000     1.556869    3.750009
  areaage1   |   3.846787   .6544836     7.92   0.000     2.755984    5.369324
------------------------------------------------------------------
```

To compare these results with the model including interaction you produced earlier, type:

**xi: logistic mf i.agebin*i.area**

```
   i.agebin   _Iagebin_0-1 (naturally coded; _Iagebin_0 omitted)
   i.area     _Iarea_0-1 (naturally coded; _Iforest_0 omitted)
   i.age~n*i.area  _IageXarea_# (coded as above)
Logistic regression                                Number of obs   =       1302
LR chi2(3)      =     300.08
Prob > chi2     =     0.0000
Log likelihood  = -706.99054                       Pseudo R2       =     0.1751
------------------------------------------------------------------
mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
   _Iagebin_1 |    5.80094   1.236674     8.25   0.000     3.819749    8.809715
   _Iarea_1   |   2.416252   .5418676     3.93   0.000     1.556869    3.750009
   _IageXarea |   1.592047   .4481527     1.65   0.099     .9169535    2.764169
------------------------------------------------------------------
```

Note that the models are identical, the log-likelihoods are equal (-706.99). In using the new variable areaage* you produced a model that directly gives the stratum-specific odds ratios for area by each age-group 5-9 years and 20+ years, as well as the 95% confidence intervals for these odds ratios.

For models where you reparameterise variables in this way:

- You need to produce the standard interaction model, with the interaction parameter(s) to test the significance of the interaction with a likelihood ratio test.
- You should only do this if there is interaction and you need to present stratum-specific estimates.
- The confidence intervals for the odds ratio estimates for the stratum-specific estimates can be used, these are difficult to obtain from the standard interaction model.
4 Stata’s *lincom* command computes point estimates, standard errors, t or z statistics, p-values, and confidence intervals for linear combinations of coefficients after any regression command:

`xi: logistic mf i.agebin*i.area`

<table>
<thead>
<tr>
<th>i.agebin</th>
<th>_Iagebin_0-1</th>
<th>(naturally coded; _Iagebin_0 omitted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i.agebin*area</td>
<td><em>IageXarea</em>#</td>
<td>(coded as above)</td>
</tr>
</tbody>
</table>

Logistic regression  
Number of obs = 1302  
LR chi2(3) = 300.08  
Prob > chi2 = 0.0000  
Log likelihood = -706.99054  
Pseudo R2 = 0.1751

| mf | Odds Ratio   | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----|--------------|-----------|------|-----|----------------------|
| _Iagebin_1 | 5.80094 | 1.236674 | 8.25 | 0.000 | 3.819749 | 8.809715 |
| area | 2.416252 | .5418676 | 3.93 | 0.000 | 1.556869 | 3.750009 |
| _IageXarea-1 | 1.592047 | .4481527 | 1.65 | 0.099 | .9169535 | 2.764169 |

The second row `area` refers to the main effect of `area`, that is the effect of forest compared to the savannah at the baseline of `agebin`, i.e. the odds ratio for microfilarial infection in the forest area compared with savannah in the 5-19 year age-group is 2.42.

For the age-stratum-specific effect on infection of exposure to forest (compared with savannah) in the age-group 20+ years, we need to combine the main effects of `area` and the interaction between `agebin` and `area`:

Forest, 20+ years: OR = 2.41 x 1.59 = 3.83

Stata’s *lincom* command does this for us, but with the additional benefit of computing 95% confidence intervals. We supply the command with the indicator variables that Stata has created for the main effects of `area` and the interaction between `agebin` and `area`:

`lincom _Iarea_1 + _IageXare_1_1`

(1) [mf]_Iarea_1 + [mf]_IageXare_1_1 = 0

| mf | Odds Ratio   | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----|--------------|-----------|------|-----|----------------------|
| (1) | 3.846787 | .6544836 | 7.92 | 0.000 | 2.755984 | 5.369324 |

In this example, both area and age are coded as binary variables. If either or both had more than two levels, Stata would create a correspondingly higher number of indicator variables. Any of these could be combined using *lincom* to compute stratum-specific odds ratios for the multiple strata.
Presenting stratum–specific estimates

If there is significant interaction, stratum-specific estimates for the effects of exposure on outcome should be presented. Here we use Stata’s `lincom` command to compute stratum-specific odds ratios after fitting a single model that includes an interaction term for age and area, plus two other variables (sex and presence of eye lesions):

```stata
xi: logistic mf i.agebin*i.area sex lesions
```

The model includes:
- `i.agebin`: age groups (5–19 years and 20+ years)
- `i.area`: geographic areas (Savannah and Forest)
- `sex`: gender (Male and Female)
- `lesions`: presence of eye lesions

The logistic regression results are as follows:

```
Logistic regression                               Number of obs   =     1302
LR chi2(5)      =     344.19
Prob > chi2     =     0.0000
Log likelihood = -684.93583                       Pseudo R2       =     0.2008
------------------------------------------------------------------
mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
   _Iagebin_1 |   5.280061   1.152462     7.62   0.000      3.44231    8.098935
   _Iarea_1 |   2.422396   .5481036     3.91   0.000     1.554705    3.774351
   _IageXare_1 |   1.506125   .4293478     1.44   0.151      .861412    2.633366
   sex |   .5928692   .0805139    -3.85   0.000     .4543206    .7736692
  lesions |   3.08199    .8173016    -4.24   0.000     1.832754    5.182726
------------------------------------------------------------------
```

We can compute adjusted odds ratios for specific strata using `lincom` commands:

```stata
lincom _Iarea_1 + _IageXare_1_1
```

```
(1) | 3.648431   .6327566     7.46   0.000     2.597049    5.125452
```

```stata
lincom _Iagebin_1 + _IageXare_1_1
```

```
(1) | 7.952433   1.508298    10.93   0.000     5.483488   11.53302
```

All of the logistic regression results can be presented in a single table as follows:

<table>
<thead>
<tr>
<th>Odds of microfilarial infection by area among different age groups:</th>
<th>Adjusted odds ratio (95% confidence interval)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 – 19 years Savannah</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>5 – 19 years Forest</td>
<td>2.42 (1.55, 3.77)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>20+ years Savannah</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>20+ years Forest</td>
<td>3.65 (2.60, 5.12)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Main effect of age in savannah 5 – 19 years</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Main effect of age in forest 5 – 19 years</td>
<td>5.28 (3.44, 8.10)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Main effect of age in forest 20+ years</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>20+ years</td>
<td>7.95 (5.48, 11.53)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Odds of microfilarial infection by other variables in model:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex Male</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Sex Female</td>
<td>0.59 (0.45, 0.77)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Presence of eye lesions No</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Presence of eye lesions Yes</td>
<td>3.08 (1.83, 5.18)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
**Key points**

1. Interaction terms should be examined and tested using the likelihood ratio test.

2. Stratum-specific estimates can be obtained by multiplying the relevant parameter estimates, by restricting the analysis to one stratum (category), or by creating new indicator variables which directly give stratum-specific estimates.

3. To include an interaction term in a logistic model you should place an asterisk (*) between the two interacting variables.

**Review exercise**

Now try to carry out the same analyses on your own. For this exercise you should use the *mwanza* dataset, a case-control study of HIV infection. The solutions are given in Section 4.

1. Using the *mwanza* data, investigate whether there is interaction between the effects of education and age. (You may need to combine categories for each variable.)

2. Assume there is interaction, and derive the appropriate age-specific odds ratios in whichever way you prefer.
Practical 9

Logistic regression 3

Aim

To learn how to include quantitative exposures in a logistic model and decide whether a linear trend is appropriate.

Learning objectives

By the end of this session you will be able to:

- include a quantitative exposure in a logistic regression model
- decide whether a linear effect or categorical effect is more appropriate
- interpret the parameter estimates for a linear effect.

Planning your session

To complete the work in this session:

- You should have worked through Session 9 of Statistical Methods in Epidemiology
- The specific dataset for this session is oncho.dta
- You must be familiar with using Stata and know the basic commands.

Getting started

Access Stata in the usual way.

In the Stata command window, change directory to the directory that you have created for your workbook practicals:

    e.g. cd c:\sme

Then create a log for your output, i.e. a file that will store all commands and output from the analysis you do in this session:

    e.g. log using smeprac9.log

Note: The log file will automatically be sent to the directory where your data files are stored, if you have specified this in the cd command.
Remember:

1. All the commands and variables you need to type in Stata are written in bold and on a separate line. The variable names and commands are also in bold when they appear in the text.
2. Stata is case-sensitive and all commands should be typed in lower case.
3. In Stata output P=0.000 indicates a P-value of P<0.001.

Note:
All commands and output are for Stata version 10.0. Most commands and output for Stata 10.0 are identical to those in Stata 9.0, and are compatible with later versions of Stata.

In this session you the models you instructed to produce are all on an odds ratio scale using the command logistic. In addition you can obtain the model estimates on a log scale using the logit command.

Reading in the dataset and identifying relevant variables

In this practical session you will again use the dataset from the prevalence study of onchocerciasis in Sierra Leone.

To read in the dataset, type:

```
use oncho
```

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

```
help oncho
```

The variables you will be working with are:

- `mf`: is the variable name for microfilarial infection coded: 0=no, 1=yes
- `area`: is the variable name for area of residence coded: 0=savanna, 1=forest
- `agegrp`: is the variable name for age-group coded: 0=5-9, 1=10-19, 2=20-39, 3=40+ years
Quantitative exposures

In most analyses it is convenient to group continuous variables such as age. You can then obtain a parameter estimate for each level of the variable compared to a baseline level. For example, with `agegrp` you have previously obtained parameter estimates for the odds ratios relative to the youngest age-group. To review these results use the following command:

**xi: logistic mf i.agegrp**

(Remember you need to specify `xi:` to produce an indicator variable for each level of `agegrp`.)

<table>
<thead>
<tr>
<th>i.agegrp _Iagegrp_0-3 (naturally coded; _Iagegrp_0 omitted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
</tr>
<tr>
<td>Number of obs = 1302</td>
</tr>
<tr>
<td>LR chi2(3) = 258.40</td>
</tr>
<tr>
<td>Prob &gt; chi2 = 0.0000</td>
</tr>
<tr>
<td>Log likelihood = -727.83149</td>
</tr>
<tr>
<td>Pseudo R2 = 0.1508</td>
</tr>
</tbody>
</table>

| mf | Odds Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----|------------|-----------|-----|-----|----------------------|
| _Iagegrp_1 | 2.821337 | .6093942 | 4.80 | 0.000 | 1.847566 - 4.308341 |
| _Iagegrp_2 | 8.112056 | 1.612103 | 10.53 | 0.000 | 5.495005 - 11.97534 |
| _Iagegrp_3 | 16.02391 | 3.334152 | 13.33 | 0.000 | 10.65752 - 24.09246 |

Assuming a linear trend

In fact, with quantitative exposures it is possible to model a linear effect (increasing or decreasing). Such a model assumes a common odds ratio, that is the relative increase (or decrease) from one age-group to the next is the same. To model this in Stata you simply do not specify that the variable is categorical. The output then assumes the same increase for each unit increase in the variable.

To do this for `agegrp` type:

**logistic mf agegrp**

| Logistic regression |
| Number of obs = 1302     |
| LR chi2(1) = 255.58      |
| Prob > chi2 = 0.0000     |
| Log likelihood = -729.23967 |
| Pseudo R2 = 0.1491       |

| mf | Odds Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----|------------|-----------|-----|-----|----------------------|
| agegrp | 2.533891 | .1615047 | 14.587 | 0.000 | 2.236321 - 2.871056 |

Note that only one parameter is required for age when a linear effect is assumed, therefore you will have a more simple model, which should always be part of your aim.

The odds ratio estimate for the linear effect of age-group is 2.53. How do you interpret this?
This is odds ratio from one level to the next, i.e. the common odds ratio for a unit increase in `agegrp`. This depends, of course, on the way in which the age-group categories are defined, and if you are reporting a linear effect in categories you should make these definitions clear to your readers.

The odds ratios for each age-group compared to the youngest age-group (5-9 years) assuming a linear effect are shown below. The estimates for separate effects are also shown.

<table>
<thead>
<tr>
<th>Age-group (years)</th>
<th>Linear effect ORs</th>
<th>Separate ORs</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-19</td>
<td>2.53</td>
<td>2.53</td>
</tr>
<tr>
<td>20-39</td>
<td>2.53² = 6.40</td>
<td>8.11</td>
</tr>
<tr>
<td>40+</td>
<td>2.53³ = 16.19</td>
<td>16.02</td>
</tr>
</tbody>
</table>

The odds ratio estimates of the linear effect and the separate effects are similar. If you can assume a linear trend, then the model is simpler. However, you must first formally assess whether the separate age-group effects provide a better model for the data, i.e. test whether there is departure from a linear trend in the separate effects of `agegrp`.

### Testing the linear assumption

Formally this is called a test of ‘departure from linear trend’. To test for departure you compare the model assuming a linear trend (OR = 2.53) to the model with separate age-group effects. This comparison is of course made with a likelihood ratio test.

You must first fit the model with most parameters, i.e. the one which models the separate effects, save the log likelihood of this model, fit the model that assumes a linear effect, then compare the two log likelihoods.

To do this, type:

```plaintext
xi: logistic mf i.agegrp
est store A
xi: logistic mf agegrp
est store B
lrtest A B
```

Likelihood-ratio test

<table>
<thead>
<tr>
<th>LR chi2(2) =</th>
<th>2.82</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob &gt; chi2  =</td>
<td>0.2446</td>
</tr>
</tbody>
</table>

The null hypothesis of this test is that the association between age-group and microfilarial infection is linear, or, more formally, that there is no difference in the goodness of fit of the two models assuming linear trend and estimating separate effects for each category. The result of this test is P=0.24. Therefore you can say that including separate effect for each age-group does not significantly improve the model and the linear effect sufficiently describes the data. In other words you can assume a linear trend in the categories which gives a more simple model.
Key points

- For quantitative exposures that have been grouped into categories a linear effect is preferable to separate effects for each category, but only if modelling with separate effects does not significantly improve the fit of the model.

- The estimate for a linear effect can be interpreted as the OR (or increase in log odds) for a unit increase in the variable. A unit increase may be an increase of one category, for a grouped categorical variable, or an increase of one unit if the variable is on its original scale (e.g. age in years).

Review exercise

Now try to carry out the same analysis on your own. For this exercise you should use the mwanza dataset which refers to a case-control study of HIV infection. The solutions are given in Section 4.

1. Using the mwanza data, produce a table of HIV infection and number of injections in the past year.

   inj is the variable name for name for injections in the past year coded: 1=none, 2=1, 3=2-4, 4=5-9, 5=10+, 9=missing

2. Use tabodds to calculate the odds of infection for each level of number of injections.

3. Produce a logistic model that assumes a linear effect of number of injections in the past year (remember to exclude the missing category).

   What is the common OR ratio from the model?

4. Use a likelihood ratio test to test for departure from a linear trend.
Matched case-control studies

**Aim**

To learn how to perform a classical analysis of matched case-control studies using Stata.

**Objectives**

By the end of this session you will be able to:

- summarise the results of pair-matched case-control studies in a table
- estimate the matched odds ratio
- test for effect modification (interaction).

**Planning your session**

To complete the work in this session:

- You should have worked through Session 10 of *Statistical Methods in Epidemiology*.
- You should have copied all the datasets and files to your working directory for these practical exercises, e.g. a directory called `asme`. The datasets for this session are `diabraz.dta` and `diabraz2.dta`. You must be familiar with using Stata.

**Getting started**

Access Stata in the usual way.

In the Stata command window change directory to the directory that you have created for your workbook practicals.

`cd asme`
Then create a log for your output, that is a file which will store all the commands and output from the analysis you do in this session
e.g. **log using asmprac2.log**

Note: The log file will automatically be sent to the directory where your data files are stored if you have specified this in the **cd** command.

**Remember:**

1. All the commands and variables you need to type in Stata are written in bold in this text on a separate line. The variable names and commands are also in bold when they appear in the text.

2. Stata is case-sensitive and all commands should be typed in lower case.

3. In Stata output P=0.000 indicates a P-value of P<0.001.

---

### 1. **Reading in the dataset and identifying relevant variables**

In this practical session you will use data from a matched case-control study of risk factors for infant mortality due to diarrhoea in Brazil. There are 2 datasets, the first has only 1 control per case, the second the complete case sets with 2 controls per case.

To illustrate the analysis for pair matched data you will first use the dataset with 1 control per case.

To read in the dataset, type:

**use diabraz**

You will see all the variables for the dataset listed in the Variables window.

For a description of the study and variables in the dataset, type:

**help diabraz**

[Note: This help file refers to both datasets for this study.]

In this analysis you will look at the effect of breast feeding, water source and type of housing on infant mortality due to diarrhoea. Controls were matched on age to the cases so we cannot assess the effect of age on infant mortality due to diarrhoea. We can however assess whether there is interaction between age and an exposure variable.
case  the variable name for case or control coded: 0=control, 1=case
agegrp2  the variable name for the matching variable age-group coded: 1=0-2 months, 2=3-5 months, 3=6-11 months
bf  the variable name for feeding with or without breast feeding. coded: 1= feeding including breast feeding, 2 = other types of feeding only
milkgp  the variable name for type of feeding mode. coded: 1 = breast feeding only; 2 = breast feeding + other types of feeding; 3 = other types of feeding only
wat2  the variable name for type of water supply coded: 1=piped water in house 2=no access to piped water
house  the variable name for type of housing coded: 1=regular building, 2=shack
bwt  the variable name for birth weight groups (in kg) coded: 1=min/2.50, 2=2.50-2.99, 3=3.00-3.49, 4=3.50+
bwtgp  the variable name for birth weight in 2 groups (in kg) coded: 1=greater or equal 3.00, 2=less than 3.00

Use commands `describe`, `summarise` and `list` to examine the data.
To produce frequency distributions use the `tab` command.

2. **A single binary exposure**

You will now examine the effect of water supply on infant mortality. For paired matched data we can produce a matched table. Type:

```
match case wat2 pair
```

<table>
<thead>
<tr>
<th>0 wat2</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+----------------------</td>
<td>-----------</td>
<td>---</td>
<td>-------</td>
</tr>
<tr>
<td>1 wat2</td>
<td>1</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>19</td>
<td>86</td>
</tr>
</tbody>
</table>

EPM202 Statistical methods in epidemiology
We can also obtain the corresponding OR estimate and hypothesis test using the `mhodds` command. Type:

```
mhodds case wat2 pair, c(2,1)
```

The option `c(2,1)` tells Stata to compare category 2 of the variable `wat2` to category 1 of `wat2`. If the option `by(pair)` is specified with the `mhodds` command, rather than including pair before the comma in the command, Stata will try to produce an OR for each individual pair as well as an overall estimate. This output is not useful and so the option `by()` is unnecessary with the `mhodds` command.

Mantel-Haenszel estimate of the odds ratio
Comparing `wat2==2` vs `wat2==1`, controlling for pair

Note: only 13 of the 86 strata formed in this analysis contribute information about the effect of the explanatory variable

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.500000</td>
<td>6.23</td>
<td>0.0126</td>
<td>1.219099 24.813414</td>
</tr>
</tbody>
</table>

This shows infants with no access to water are at 5.5 times the odds of mortality than those who have access to water in their house/plot.

Let's see what happens if we ignore matching in the analysis.

```
mhodds case wat2, c(2,1)
```

Maximum likelihood estimate of the odds ratio
Comparing `wat2==2` vs `wat2==1`

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.702359</td>
<td>2.36</td>
<td>0.1247</td>
<td>0.856308 3.384327</td>
</tr>
</tbody>
</table>

Ignoring the matching in the analysis results in the odds ratio being underestimated. (OR matched = 5.50, OR unmatched = 1.70.) This is due to water supply being strongly associated with the matching factor (neighbourhood).

### 3. Interaction with a matching variable

Although we cannot examine the effect of a matching variable we can examine whether there is any evidence that a matching variable modifies the effect of the exposure of interest. To do this we need to look at the effect of the risk factor at different levels of the matching variable.
Because of the small numbers within strata it is better to group age. First group age into categories: 0-2 months and 3+ months. Type:

```
gen agenew = agegp2
```

```
recode agenew 3=2
```

Now use the following commands to produce a matched table for the effect of water supply on infant mortality at each level of agenew.

```
match case wat2 pair if agenew ==1
```

<table>
<thead>
<tr>
<th></th>
<th>0 wat2</th>
<th></th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>7</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

```
mhodds case wat2 pair if agenew==1, c(2,1)
```

Mantel-Haenszel estimate controlling for pair

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.000000</td>
<td>1.80</td>
<td>0.1797</td>
<td>0.447083 35.787573</td>
</tr>
</tbody>
</table>

```
match case wat2 pair if agenew ==2
```

<table>
<thead>
<tr>
<th></th>
<th>0 wat2</th>
<th></th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>1</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>11</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>12</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

```
mhodds case wat2 pair if agenew==2, c(2,1)
```

Mantel-Haenszel estimate controlling for pair

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.000000</td>
<td>4.50</td>
<td>0.0339</td>
<td>0.861242 56.894563</td>
</tr>
</tbody>
</table>

To test for heterogeneity of the odds ratios within the separate strata you should make a table of the discordant pairs.

<table>
<thead>
<tr>
<th>Age group</th>
<th>0-2 months</th>
<th>3+ months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case + control -</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Case – control +</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Water supply: -ve supply in house or plot, +ve no water supply

To obtain the corresponding chi-squared test for this table use the immediate tabulate command.

```
tabi 4 7 \ 11, chi
```

<table>
<thead>
<tr>
<th>row</th>
<th>col</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>Total</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

Pearson chi2(1) = 0.1330   Pr = 0.715

P=0.7, showing no evidence that the effect of water supply on infant mortality is modified by age group (0-2 months, 3+ months).

---

### 4. More than 1 control per case

The classical analysis of matched case-control studies with more than 1 control per case is similar to that for paired studies. However, in Stata it is not easy to produce a matched frequency table.

Read in the dataset with 2 controls per case for the Brazilian matched case-control study.

```
use diabraz2, clear
```

Because there is more than 1 control per case, the variable set (as in case set) is used to match the data.

Now assess the effect of breast feeding on infant mortality due to diarrhoea.

```
mhodds case bf set, c(2,1)
```

**Mantel-Haenszel estimate controlling for set**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>chi2(1)</td>
<td>P&gt;chi2</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>4.05824</td>
<td>41.28</td>
<td>0.0000</td>
<td>2.554216 6.449748</td>
</tr>
</tbody>
</table>

Thus, infants who were not breast fed at all were at 4 times the risk of infant mortality than infants who were completely or partially breast fed.
5. **Exposures with more than 2 levels and adjusting for other risk factors**

To analyse exposures with more than 2 levels using classical matched analysis it is necessary to restrict the comparison to 2 levels at a time. (This can be done using the Stata commands above.) However, the results of this analysis will not be the same as if all the levels were included in the analysis.

In this situation, it is preferable (if not necessary) to use a regression model, specifically conditional logistic regression, which you will learn about in the following session.

Similarly, to take account of other risk factors of interest and assess potential confounding or interaction you must also use conditional logistic regression.

**Review exercise**

Now try to carry out the same analyses on your own. For this exercise you will use the two datasets `diabraz.dta` and `diabraz2.dta`.

Read in the paired matched dataset, `diabraz`.

1. What is the effect of birth weight (bwtgp) on infant mortality due to diarrhoea?

2. Is there any evidence of interaction between birth weight and age? (Use age in 2 groups 0-2 month, 3+ months.)

Now read in the dataset with 2 controls per case, `diabraz2`.

3. What is the effect of birth weight in this dataset?

4. Is there any evidence of interaction with age as defined above?
Section 4
Solutions to practical sessions
(1-4 and 7-10 only)
Solutions

You should first open a log files to save your results.

e.g. log using smeex1.log

For the first part of this exercise you need to use the Whitehall dataset. You should have read in the dataset containing the new recoded variables whitehall2.dta.

use whitehall2, clear

*Note: the command clear clears the current dataset and allows you to read in a new one.*

1. **How many men in the dataset died of CHD?**

   ```
   tab chd
   ```

<table>
<thead>
<tr>
<th>death from</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>chd</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1523</td>
<td>90.82</td>
<td>90.82</td>
</tr>
<tr>
<td>1</td>
<td>154</td>
<td>9.18</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>1677</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

154 men died of CHD

2. **What is the risk of death from CHD in each smoking category of smok2?**

   ```
   tab smok2 chd, row
   ```

<table>
<thead>
<tr>
<th>smok2</th>
<th>death from chd</th>
<th>0</th>
<th>l</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>897</td>
<td>66</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td></td>
<td>93.15</td>
<td>6.85</td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>626</td>
<td>88</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87.68</td>
<td>12.32</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1523</td>
<td>154</td>
<td>1677</td>
</tr>
</tbody>
</table>

   |          | 90.82 | 9.18 | 100.00|

The risk of death from CHD is 6.85% (0.0685) in current non-smokers and 12.32% (0.123) in current smokers.
3. **What is the risk ratio of current smokers to non-smokers?**

The risk ratio is calculated as the ratio of the two risks

$$RR = \frac{12.32}{6.85} = 1.80$$

or you can use the `cs` command:

```
cs chd smok2
```

<table>
<thead>
<tr>
<th>smok2</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Unexposed</td>
</tr>
<tr>
<td>Cases</td>
<td>88</td>
<td>66</td>
</tr>
<tr>
<td>Noncases</td>
<td>626</td>
<td>897</td>
</tr>
<tr>
<td>Total</td>
<td>714</td>
<td>963</td>
</tr>
<tr>
<td>Risk</td>
<td>.1232493</td>
<td>.0685358</td>
</tr>
<tr>
<td></td>
<td>Point estimate</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>Risk difference</td>
<td>.0547135</td>
<td>.0257992    .0836277</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>1.798319</td>
<td>1.326753    2.437494</td>
</tr>
<tr>
<td>Attr. frac. ex.</td>
<td>.4439252</td>
<td>.2462803    .5897425</td>
</tr>
<tr>
<td>Attr. frac. pop</td>
<td>.2536716</td>
<td>--    --    --</td>
</tr>
</tbody>
</table>

**chi2(1) = 14.72  Pr>chi2 = 0.0001**

The risk ratio for smokers to non-smokers is 1.798.

4. **Use the stset command to define follow-up details for the outcome CHD**

**Note:** In the `stset` command, the options after the comma do not have to be in a particular order.

```
stset timeout, fail(chd) id(id) origin(timein) scale(365.25)
```

```
id:  id
failure event:  chd != 0 & chd < .
obs. time interval:  (timeout[_n-1], timeout]
exit on or before:  failure
t for analysis:  (time-origin)/365.25
origin:  time timein
```

```
1677  total obs.
0  exclusions
---------------------------------------
1677  obs. remaining, representing
1677 subjects
154 failures in single failure-per-subject data
27605.37 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 19.38123
```
5. **Use `strate` to obtain the CHD mortality rates for current smokers and non-smokers**

```bash
strate smok2

failure _d: chd
analysis time _t: (timeout-origin)/365.25
origin: time timein
id: id
```

Estimated rates and lower/upper bounds of 95% confidence intervals (1677 records included in the analysis)

```
<table>
<thead>
<tr>
<th>smok2</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>66</td>
<td>1.6e+04</td>
<td>0.0040336</td>
<td>0.0031689</td>
<td>0.0051341</td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td>1.1e+04</td>
<td>0.0078273</td>
<td>0.0063514</td>
<td>0.0096460</td>
</tr>
</tbody>
</table>
```

The rate of deaths from CHD for non-smokers and ex-smokers combined is 0.0040 per person-year.

The rate of deaths from CHD for current smokers is 0.0078 per person-year.

6. **What is the rate of deaths due to CHD per 1000 person-years for current smokers and non-smokers?**

```bash
strate smok2, per(1000)

failure _d: chd
analysis time _t: (timeout-origin)/365.25
origin: time timein
id: id
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals (1677 records included in the analysis)

```
<table>
<thead>
<tr>
<th>smok2</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>66</td>
<td>16.3626</td>
<td>4.0336</td>
<td>3.1689</td>
<td>5.1341</td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td>11.2427</td>
<td>7.8273</td>
<td>6.3514</td>
<td>9.6460</td>
</tr>
</tbody>
</table>
```

The rate of death for non-smokers is 4.03 per 1000 person-years.

The rate of death for current smokers is 7.83 per 1000 person-years.
7. **Use stmh to obtain the rate ratio of CHD death rates for smokers to non-smokers**

stmh smok2, c(1,0)

failure _d:  chd
analysis time _t:  (timeout-origin)/365.25
origin:  time timein
id:  id

Maximum likelihood estimate of the rate ratio comparing smok2==1 vs. smok2==0

RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.941</td>
<td>17.19</td>
<td>0.0000</td>
<td>1.410, 2.670</td>
</tr>
</tbody>
</table>

The rate ratio for current smokers to non-smokers is 1.94. So the rate of CHD deaths in smokers is almost twice that of non-smokers.

If you want to save your data, type:

save whitehall2, replace

8. **Read in the Mwanza data**

use mwanza, clear

9. **How many women had ever used a condom?**

   tab usedc

<table>
<thead>
<tr>
<th>Ever used a condom</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>738</td>
<td>96.72</td>
<td>96.72</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>3.01</td>
<td>99.74</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.26</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 763   | 100.00 |

The codes for ever used a condom are 1=no, 2=yes, and 9=missing.

So, 23 women had ever used a condom.
10. **What proportion of cases had ever used a condom?**

```stata
tab case usedc, row
```

<table>
<thead>
<tr>
<th>Case/control</th>
<th>Ever used a condom</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>Total</td>
</tr>
<tr>
<td>0</td>
<td>558</td>
<td>16</td>
<td>0</td>
<td>574</td>
</tr>
<tr>
<td></td>
<td>97.21</td>
<td>2.79</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>1</td>
<td>180</td>
<td>7</td>
<td>2</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>95.24</td>
<td>3.70</td>
<td>1.06</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>738</td>
<td>23</td>
<td>2</td>
<td>763</td>
</tr>
<tr>
<td></td>
<td>96.72</td>
<td>3.01</td>
<td>0.26</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The codes for ever used a condom are 1=no, 2=yes, and 9=missing. So of all the cases (including the missing data) 7/189 = 0.037 (3.7%) had ever used a condom.

11. **Use *mhodds* to obtain the odds ratio for condom use and HIV infection**

```stata
recode usedc 9=.  
mhodds case usedc, c(2,1)
```

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.356250</td>
<td>0.44</td>
<td>0.5076</td>
<td>0.548775 3.351854</td>
</tr>
</tbody>
</table>

The odds ratio for condom use and HIV infection is 1.36. However, this association is not significant (P=0.51) and from the 95% confidence interval you can say the true odds ratio is likely to be between 0.55 and 3.35. So it’s possible that the odds of HIV infection could be higher in women who do not use condoms.

Remember to close your log file:

```stata
log close
```
Summary

In this exercise you have

- Investigated the risk of death from CHD for smokers and non-smokers, and calculated a risk ratio compare risks for the 2 groups
- Used the Stata command `stset` to define follow-up information for the “whitehall2” dataset
- Used the Stata command `strate` to obtain CHD mortality rates for smokers and non-smokers
- Use the Stata command `stmh` to obtain the rate ratio of CHD death rates comparing smokers with non-smokers
- Revised the use of the Stata command `mhodds` to calculate an Odds Ratio which compares odds of HIV infection amongst women ever using, and not using, a condom.
Solutions

You should first open a log file to save your results.
e.g. log using smeex2.log

Read in the original Whitehall dataset again
use whitehal, clear

1. Defining CHD mortality as the outcome

In order to examine CHD mortality you need to redefine CHD mortality. First we need to redefine stset:

\texttt{stset timeout, fail(chd) origin(timein) id(id) scale(365.25)}

\begin{verbatim}
    id: id
    failure event: chd != 0 & chd < .
    obs. time interval: [timeout[=_n-1], timeout]
    exit on or before: failure
    t for analysis: (time-origin)/365.25
    origin: time timein

    1677  total obs.
    0  exclusions
    -------
    1677  obs. remaining, representing
    1677 subjects
    154 failures in single failure-per-subject data
    27605.37 total analysis time at risk, at risk from t = 0
    earliest observed entry t = 0
    last observed exit t = 19.38123
\end{verbatim}

There are 1677 subjects in the dataset. The follow-up time starts at entry, hence the minimum is 0 years, and lasts up to a maximum of 19.4 years. There were 154 ‘failures’, i.e. deaths from CHD. Use the tab command to check that the number of ‘failures’ tallies with this.

\texttt{tab chd}

\begin{verbatim}
   death from | chd  | Freq. | Percent | Cum.
   -------- | ---- | ----- | ------ | ----
   0              | 1,523 | 90.82 | 90.82  |
   1              | 154  | 9.18  | 100.00 |
   -------
   Total          | 1,677| 100.00|
\end{verbatim}

The table confirms that the number of deaths from CHD was 154.
2. **Recode age at entry**

```stata
egen agecat=cut(agein), at(40,45,50,55,60,65,70) label
strat agecat
```

Estimated rates and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)

<table>
<thead>
<tr>
<th>agecat D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>11</td>
<td>4.9e+03</td>
<td>0.0022364</td>
<td>0.0012385</td>
</tr>
<tr>
<td>45</td>
<td>16</td>
<td>7.9e+03</td>
<td>0.0020369</td>
<td>0.0012479</td>
</tr>
<tr>
<td>50</td>
<td>34</td>
<td>6.1e+03</td>
<td>0.0056108</td>
<td>0.0040091</td>
</tr>
<tr>
<td>55</td>
<td>50</td>
<td>5.3e+03</td>
<td>0.0094661</td>
<td>0.0071745</td>
</tr>
<tr>
<td>60</td>
<td>32</td>
<td>3.1e+03</td>
<td>0.0103398</td>
<td>0.0073120</td>
</tr>
<tr>
<td>65</td>
<td>11</td>
<td>395.2472</td>
<td>0.0278307</td>
<td>0.0154126</td>
</tr>
</tbody>
</table>

3. **CHD mortality rates and rate ratio for grade of employment**

To examine the mortality rates in each grade of employment, use `strat`. If you use the option `per(1000)` the rates are given per 1000 person-years.

```stata
strat grade, per(1000)
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)

<table>
<thead>
<tr>
<th>grade</th>
<th>D</th>
<th>Y</th>
<th>Rate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>20.3398</td>
<td>4.4248</td>
<td>3.5989</td>
<td>5.4403</td>
</tr>
<tr>
<td>2</td>
<td>64</td>
<td>7.2656</td>
<td>8.8087</td>
<td>6.8946</td>
<td>11.2541</td>
</tr>
</tbody>
</table>

To obtain a rate ratio for the effect of low grade compared to high grade use `stmh`.

```stata
stmh grade, c(2,1)
```

Maximum likelihood estimate of the rate ratio
comparing grade==2 vs. grade==1

RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.991</td>
<td>18.44</td>
<td>0.0000</td>
<td>1.445 2.743</td>
</tr>
</tbody>
</table>

The rate of CHD mortality is almost twice as high in low-grade workers compared to high-grade workers (P<0.0001).
4. **Effect of grade by age at entry**

```stata
stmh grade, by(agecat) c(2,1)
```

Maximum likelihood estimate of the rate ratio comparing grade==2 vs. grade==1 by agecat

RR estimate, and lower and upper 95% confidence limits

<table>
<thead>
<tr>
<th>agecat</th>
<th>RR</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-</td>
<td>2.29</td>
<td>0.61</td>
<td>8.62</td>
</tr>
<tr>
<td>45-</td>
<td>1.10</td>
<td>0.31</td>
<td>3.85</td>
</tr>
<tr>
<td>50-</td>
<td>2.33</td>
<td>1.18</td>
<td>4.61</td>
</tr>
<tr>
<td>55-</td>
<td>1.48</td>
<td>0.85</td>
<td>2.59</td>
</tr>
<tr>
<td>60-</td>
<td>0.67</td>
<td>0.33</td>
<td>1.35</td>
</tr>
<tr>
<td>65-</td>
<td>0.67</td>
<td>0.18</td>
<td>2.51</td>
</tr>
</tbody>
</table>

Overall estimate controlling for agecat

<table>
<thead>
<tr>
<th>RR</th>
<th>chi2</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.286</td>
<td>2.35</td>
<td>0.1249</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.774</td>
</tr>
</tbody>
</table>

Approx test for unequal RRs (effect modification): \( \chi^2(5) = 8.46 \), \( P>\chi^2 = 0.1326 \)

Although the stratum-specific RR estimates vary from 0.67 to 2.33, the variation is not systematic.

Notice how some rate ratios show an increased rate for low-grade compared to high-grade, and others a decreased rate. The test for effect modification is not significant. So you can conclude that there is no evidence of interaction between employment grade and age at entry.

The Mantel-Haenszel estimate is \( RR=1.29 \) (95% CI 0.93 to 1.77) indicating that the original estimate of the RR for CHD mortality, \( RR=1.99 \) (95% CI 1.45 to 2.74), was partly confounded by age at entry.

**Summary**

In this exercise you have

- Used the Stata command `stset` to define follow-up information for the “whitehal” dataset, using both age at entry into the study and date of birth as the time variables
- Used the Stata command `egen` to divide the quantititative age variable into age-group categories
- Used the Stata command `stmh` to obtain a rate ratios for CHD comparing grade of employment, both unadjusted and adjusted for age at entry into the study.
You should first open a log file to save your results.

\texttt{e.g. log using smeex3.log}

Read in the Trinidad dataset.

\texttt{use trinidad, clear}

\section*{1. Producing a lifetable for current smokers}

\texttt{ltable years death if current==1}

\begin{tabular}{llllllll}
\hline
Interval & Total & Deaths & Lost & Survival & Std. Error & [95\% Conf. Int.] \\
\hline
0 & 1 & 109 & 4 & 0 & 0.9633 & 0.0180 & 0.9052 & 0.9861 \\
1 & 2 & 105 & 5 & 0 & 0.9174 & 0.0264 & 0.8473 & 0.9562 \\
2 & 3 & 100 & 4 & 0 & 0.8807 & 0.0310 & 0.8035 & 0.9289 \\
3 & 4 & 96 & 5 & 0 & 0.8349 & 0.0356 & 0.7508 & 0.8926 \\
4 & 5 & 91 & 7 & 1 & 0.7703 & 0.0403 & 0.6793 & 0.8385 \\
5 & 6 & 83 & 2 & 4 & 0.7513 & 0.0415 & 0.6586 & 0.8221 \\
6 & 7 & 77 & 2 & 5 & 0.7311 & 0.0428 & 0.6365 & 0.8048 \\
7 & 8 & 70 & 7 & 12 & 0.6511 & 0.0476 & 0.5493 & 0.7355 \\
8 & 9 & 51 & 3 & 27 & 0.5990 & 0.0524 & 0.4886 & 0.6931 \\
9 & 10 & 21 & 1 & 20 & 0.5446 & 0.0705 & 0.3975 & 0.6701 \\
\hline
\end{tabular}

The 5\textsuperscript{th} year occurs in interval 4 to 5. There were 7 deaths from any cause in this year. The probability of survival for current smokers at the end of this year is 0.77.
2. **Comparison of current smokers to current non-smokers**

**Itable years death, by(current)**

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg.</th>
<th>Total</th>
<th>Deaths</th>
<th>Lost</th>
<th>Survival</th>
<th>Std. Error</th>
<th>[95% Conf. Int.]</th>
</tr>
</thead>
<tbody>
<tr>
<td>current = 0</td>
<td>0</td>
<td>208</td>
<td>12</td>
<td>0</td>
<td>0.9423</td>
<td>0.0162</td>
<td>0.9006</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>196</td>
<td>6</td>
<td>0</td>
<td>0.9135</td>
<td>0.0195</td>
<td>0.8862</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>190</td>
<td>3</td>
<td>1</td>
<td>0.8990</td>
<td>0.0209</td>
<td>0.8493</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>186</td>
<td>9</td>
<td>1</td>
<td>0.8554</td>
<td>0.0244</td>
<td>0.7997</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>176</td>
<td>5</td>
<td>0</td>
<td>0.8311</td>
<td>0.0260</td>
<td>0.7727</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>171</td>
<td>2</td>
<td>18</td>
<td>0.8208</td>
<td>0.0267</td>
<td>0.7613</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>151</td>
<td>3</td>
<td>24</td>
<td>0.8031</td>
<td>0.0280</td>
<td>0.7412</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>124</td>
<td>6</td>
<td>17</td>
<td>0.7614</td>
<td>0.0313</td>
<td>0.6932</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>101</td>
<td>2</td>
<td>58</td>
<td>0.7402</td>
<td>0.0338</td>
<td>0.6669</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>41</td>
<td>0</td>
<td>41</td>
<td>0.7402</td>
<td>0.0338</td>
<td>0.6669</td>
</tr>
<tr>
<td>current = 1</td>
<td>0</td>
<td>109</td>
<td>4</td>
<td>0</td>
<td>0.9633</td>
<td>0.0180</td>
<td>0.9052</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>105</td>
<td>5</td>
<td>0</td>
<td>0.9174</td>
<td>0.0264</td>
<td>0.8473</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>100</td>
<td>4</td>
<td>0</td>
<td>0.8807</td>
<td>0.0310</td>
<td>0.8035</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>96</td>
<td>5</td>
<td>0</td>
<td>0.8349</td>
<td>0.0356</td>
<td>0.7508</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>91</td>
<td>7</td>
<td>1</td>
<td>0.7703</td>
<td>0.0403</td>
<td>0.6793</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>83</td>
<td>2</td>
<td>4</td>
<td>0.7513</td>
<td>0.0415</td>
<td>0.6586</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>77</td>
<td>2</td>
<td>5</td>
<td>0.7311</td>
<td>0.0428</td>
<td>0.6365</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>70</td>
<td>7</td>
<td>12</td>
<td>0.6511</td>
<td>0.0476</td>
<td>0.5493</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>51</td>
<td>3</td>
<td>27</td>
<td>0.5990</td>
<td>0.0524</td>
<td>0.4886</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>21</td>
<td>1</td>
<td>20</td>
<td>0.5446</td>
<td>0.0705</td>
<td>0.3975</td>
</tr>
</tbody>
</table>

After the first two intervals, the survival probability for current smokers is always lower than for non-smokers throughout the follow-up period. At the end of follow-up the cumulative survival for current non-smokers is 0.74 compared to that for current smokers, which is 0.54.

3. **Kaplan-Meier survival curves by smoking status**

To produce a Kaplan-Meier plot you should first set the follow-up time information using `stset`. You must be careful to specify when individuals first become at risk using the `origin` option and the start of the follow-up using the `enter` option.

```
stset timeout, fail(death) enter(timein) origin(timein) id(id) scale(365.25)
```

```
  id: id
  failure event: death != 0 & death < .
obs. time interval: [timeout[_n-1], timeout]
enter on or after: time timein
exit on or before: failure
t for analysis: (time-origin)/365.25
origin: time timein
```

| 318 total obs. |
| 0 exclusions |

| 318 obs. remaining, representing |
| 318 subjects |
| 88 failures in single failure-per-subject data |

```
2204.539 total analysis time at risk, at risk from t = 0
earliest observed entry t = 0
last observed exit t = 9.798768
```

There were 88 deaths during follow-up.
4. **Logrank test to compare survival curves of current smokers and non-smokers**

```stata
sts test current
failure _d:  death
analysis time _t:  (timeout-origin)/365.25
origin:  time timein
enter on or after:  time timein
id:  id

Log-rank test for equality of survivor functions

<table>
<thead>
<tr>
<th>Events</th>
<th>current</th>
<th>observed</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>48</td>
<td>57.86</td>
<td>57.86</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>30.14</td>
<td>30.14</td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>88.00</td>
<td>88.00</td>
</tr>
</tbody>
</table>

chi2(1) = 4.91
Pr>chi2 = 0.0267
```

The logrank test shows that the survival probability for current smokers is significantly lower than that for current non-smokers (P=0.03). Note that you cannot see whether the survival probability is lower or higher from this test; we have established this by looking at the survival patterns above.

Remember to close your log file.

**log close**
Summary

In this exercise you have

- Used the Stata command `ltable` to produce lifetable estimates for current smokers in the “Trinidad” dataset
- Used the Stata command `sts graph` to produce a Kaplan-Meier survival curves according to smoking status using the “Trinidad” dataset
- Used the Stata command `sts test` to perform the log rank test comparing survival for smokers and non-smokers in the “Trinidad” dataset.
You should first open a log file to save your results.

* e.g. log using smeex4.log

Read in the Mwanza dataset.

* use mwanza, clear

---

1. **Does religion confound the association between schooling and HIV infection?**

```plaintext
tab rel

+----------+-------+---------+------|
1 | 48 | 6.29 | 6.29 |
2 | 321 | 42.07 | 48.36 |
3 | 205 | 26.87 | 75.23 |
4 | 188 | 24.64 | 99.87 |
9 | 1 | 0.13 | 100.00 |
+----------+-------+---------+------|
Total | 763 | 100.00 |

Religion codes: 1=Moslem, 2=Catholic, 3=Protestant, 4=other, 9=missing.

Religion has a code 9 for missing values, you should set this to system-missing (.):

* recode rel 9=.

```plaintext
\[\text{tab case rel, chi row}\]

Case/control | 1 | 2 | 3 | 4 | Total
+------------+---+---+---+---+------|
0 | 28 | 228 | 150 | 167 | 573
| 4.89 | 39.79 | 26.18 | 29.14 | 100.00 |
1 | 20 | 93 | 55 | 21 | 189
| 10.58 | 49.21 | 29.10 | 11.11 | 100.00 |
+------------+---+---+---+---+------|
Total | 48 | 321 | 205 | 188 | 762
| 6.30 | 42.13 | 26.90 | 24.67 | 100.00 |

Pearson chi2(3) = 29.4949 Pr = 0.000
```

Note: row percentages are used, not column percentages, because cases and controls had different chances of selection.

Religion is associated with HIV infection. Other religions (code 4) are clearly under-represented among cases relative to controls.
You could also calculate odds ratios for having HIV infection comparing each religious group in turn. With Moslems the resulting ORs are 0.57, 0.51, and 0.18, respectively.

To see whether potential confounder is associated with the exposure look at the relationship among controls.

```
tab ed2 rel if case==0, chi col
```

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ed2</td>
<td>13</td>
<td>80</td>
<td>42</td>
<td>128</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td>46.43</td>
<td>35.09</td>
<td>28.00</td>
<td>76.65</td>
<td>45.90</td>
</tr>
<tr>
<td></td>
<td>53.57</td>
<td>64.91</td>
<td>72.00</td>
<td>23.35</td>
<td>54.10</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>228</td>
<td>150</td>
<td>167</td>
<td>573</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Pearson \( \chi^2(3) = 93.6698 \)  \( Pr = 0.000 \)

Therefore, religion is associated with education (\( P<0.001 \)). People in the 'other' religions category are least likely to have education, and protestants are most likely.

To confine the Mantel-Haenszel OR to those with known religion, we can exclude the missing religion.

```
mhodds case ed2 if rel!=., by(rel) c(2,1)
```

Note: although the option c(2,1) is not strictly necessary here since the default option in STATA is to compare the higher-numbered code category with the lower one, it is good practice to include it to remind you which category you want to be the reference group before entering the command.

```
Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs ed2==1
by rel

<table>
<thead>
<tr>
<th>rel</th>
<th>Odds ratio</th>
<th>( \chi^2(1) )</th>
<th>P&gt;( \chi^2 )</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.022222</td>
<td>1.29</td>
<td>0.2562</td>
<td>0.584714  6.993816</td>
</tr>
<tr>
<td>2</td>
<td>2.252252</td>
<td>7.69</td>
<td>0.0056</td>
<td>1.248565  4.062776</td>
</tr>
<tr>
<td>3</td>
<td>1.393519</td>
<td>0.79</td>
<td>0.3745</td>
<td>0.667751  2.908110</td>
</tr>
<tr>
<td>4</td>
<td>2.019724</td>
<td>2.15</td>
<td>0.1425</td>
<td>0.774144  5.269413</td>
</tr>
</tbody>
</table>

Mantel-Haenszel estimate controlling for rel

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>( \chi^2(1) )</th>
<th>P&gt;( \chi^2 )</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.914248</td>
<td>10.89</td>
<td>0.0010</td>
<td>1.292931  2.834138</td>
</tr>
</tbody>
</table>

Test of homogeneity of ORs (approx): \( \chi^2(3) = 1.03 \)
\( Pr>\chi^2 = 0.7931 \)
```

The stratum-specific ratios look fairly similar (2.02, 2.25, 1.39, 2.02). The third group seems to have a lower odds ratio than the others, but the 95% confidence interval substantially overlaps that of the others. There is not much sign of interaction/effect modification, which is confirmed by the test result.
The summary adjusted odds ratio controlled for religion (1.91) is lower than the crude odds ratio of 2.42. This is what we might expect given that those individuals at lowest risk of infection were also those least likely to have gone to school.

The crude odds ratio calculated should refer to exactly the same group of individuals as the adjusted odds ratio. Therefore, in this case we should find the crude odds ratio for the group omitting the person whose religion is unknown.

\[ \text{mhodds case ed2 if rel!=., c(2,1)} \]

Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs ed2==1

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.423963</td>
<td>23.42</td>
<td>0.0000</td>
<td>1.673565 3.510826</td>
</tr>
</tbody>
</table>

One missing case won’t make much difference in a large sample, but there can be a substantial difference where there are several missing values. Therefore, it is good practice always to check that the crude and adjusted estimates apply to the same population.

2. **Test for trend on number of sexual partners**

To investigate the possibility of increasing risk of HIV infection with number of sexual partners you should carry out a test for trend.

\[ \text{tab npa} \]

<table>
<thead>
<tr>
<th>Number of</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>partners</td>
<td>ever</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Freq.</td>
<td>Percent</td>
<td>Cum.</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>26.21</td>
<td>26.21</td>
</tr>
<tr>
<td>2</td>
<td>369</td>
<td>48.36</td>
<td>74.57</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>16.12</td>
<td>90.69</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>5.64</td>
<td>96.33</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>3.67</td>
<td>100.00</td>
</tr>
</tbody>
</table>

To exclude missing values for npa you should recode them to system missing:

\[ \text{recode npa 9=}. \]
tabodds case npa

<table>
<thead>
<tr>
<th>npa</th>
<th>cases</th>
<th>controls</th>
<th>odds</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>173</td>
<td>0.15607</td>
<td>0.10404 0.23413</td>
</tr>
<tr>
<td>2</td>
<td>92</td>
<td>277</td>
<td>0.33213</td>
<td>0.26235 0.42047</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>83</td>
<td>0.48193</td>
<td>0.33048 0.70278</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>19</td>
<td>1.26316</td>
<td>0.69194 2.30592</td>
</tr>
</tbody>
</table>

Test of homogeneity (equal odds): chi2(3) = 39.64
Pr>chi2 = 0.0000

Score test for trend of odds: chi2(1) = 37.26
Pr>chi2 = 0.0000

There appears to be an increasing trend in the odds of HIV infection with an increase in number of sexual partners (although, strictly, these are not the odds because of the different selection probabilities among cases and controls in a case-control study).

The test for homogeneity, P<0.0001, gives evidence that the odds are not the same in each category of npa.

The test for trend shows strong evidence of significant increasing trend, P<0.0001.

You can obtain a common OR for this trend using mhodds.

mhodds case npa

Score test for trend of odds with npa
(The OR estimate is an approximation to the odds ratio for a one unit increase in npa)

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.885332</td>
<td>37.26</td>
<td>0.0000</td>
<td>1.538023 2.311068</td>
</tr>
</tbody>
</table>

The odds increase by 89% (times 1.89) for every increase in category of npa. There is evidence of a fairly steep increase in odds of HIV infection with increasing numbers of partners.

Summary

In this exercise you have

- Used the Stata command mhodds to calculate the Mantel Haenszel OR for the association between HIV infection and education, adjusting for religion, in the “Mwanza” dataset.
- Used the Stata command tabodds to test for a trend in the risk of HIV infection with number of sexual partners.
Practical 7

Solutions

You should first open a log file to save your results.

\texttt{e.g. log using smeex7.log}

Read in the oncho dataset.

\texttt{use oncho, clear}

1. 

\textbf{Association between microfilarial infection and sex}

To examine the association between microfilarial infection and sex first produce a table with appropriate percentages.

\texttt{tab mf sex, col chi}

<table>
<thead>
<tr>
<th>Microfil. infection</th>
<th>Sex</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>190</td>
<td>290</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.84</td>
<td>42.27</td>
<td>36.87</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>426</td>
<td>396</td>
<td>822</td>
</tr>
<tr>
<td></td>
<td></td>
<td>69.16</td>
<td>57.73</td>
<td>63.13</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>616</td>
<td>686</td>
<td>1302</td>
</tr>
</tbody>
</table>

Pearson chi²(1) = 18.2174   Pr = 0.000

You can see that 69\% of men were infected compared to 58\% of women. Sex appears to be significantly associated with microfilarial infection, \(P<0.001\).

To produce a logistic model it does not matter if you specify sex as an indicator variable or not since it is a binary variable (coded 0,1).

\texttt{logistic mf sex}

| mf | Odds Ratio | Std. Err. | z     | P>|z|  | [95\% Conf. Interval] |
|----|------------|-----------|-------|------|-------------------------|
| sex| 0.6090335  | 0.0709832 | -4.255| 0.000| 0.4846555 - 0.7653309   |

Note: \texttt{xi: logistic mf i.sex} would give the same results.

The OR estimate for sex is 0.61 (95\% CI 0.48 to 0.77). You can say that the odds of microfilarial infection for women is estimated to be 39\% lower than the odds of microfilarial infection for men. There is a strong and highly significant association between microfilarial infection and sex (\(P<0.001\)).
2. **Controlling for the effects of area and age-group**

To control the effect of sex on microfilarial infection for the possible confounding effects of area and agegrp simply include them in the logistic model.

```
xi: logistic mf i.area i.agegrp i.sex
```

Logistic regression

- Number of obs = 1302
- LR chi2(5) = 347.96
- Prob > chi2 = 0.0000

Log Likelihood = -683.05151
- Pseudo R2 = 0.2030

| mf         | Odds Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|------------|------------|-----------|-----|-----|----------------------|
| Iarea_1    | 3.073138   | 0.425981  | 8.099 | 0.000 | 2.342034    | 4.032467 |
| Iagegr_1   | 2.567357   | 0.5748449 | 4.211 | 0.000 | 1.65538    | 3.981758 |
| Iagegr_2   | 10.46237   | 2.205354  | 11.138 | 0.000 | 6.921597    | 15.81446 |
| Iagegr_3   | 17.65935   | 3.834073  | 13.225 | 0.000 | 11.53899    | 27.02598 |
| Isex_1     | 0.5591696  | 0.0758462 | -4.286 | 0.000 | 0.4286332   | 0.7294595 |

The adjusted OR estimate for sex is 0.56 (95% CI 0.43 to 0.73), very slightly lower than the crude estimate. You can say that the association between microfilarial infection and sex does not change after controlling for the other variables.

3. **Should sex be included in the model?**

To test whether sex should be included in the model you should compare a model with all three exposure to a model excluding sex using a likelihood ratio test.

```
xi: logistic mf i.agegrp i.area
est store A
xi: logistic mf i.agegrp
est store B
lrtest A B
```

- LR chi2(1) = 18.71
- Prob > chi2 = 0.0000

The likelihood ratio for sex is P<0.0001. Sex makes a large contribution to the fit of the model even after controlling for the effects of area and age.

**Summary**

In this exercise you have

- Used the `logit` command in Stata to perform logistic regression analyses of the relationship between microfilarial infection (outcome) and
  
  1. a single binary exposure variable (sex)
  2. an exposure variable with more than 2 levels (age-group)
  3. several exposure variables. All logistic regression analyses in this exercise were performed using the `logit` command in Stata.
- Performed the Likelihood Ratio Test using the Stata command `lrtest` to decide whether or not sex should be included in the model looking at risk factors for microfilarial infection.
You should first open a log file to save your results.
e.g. log using smeex8.log

Read in the mwanza dataset.
use mwanza, clear

1. **Investigate whether there is an interaction between the effects of education and age**

   The following is a suggested approach; there are many alternatives.

   First combine some levels of education and age.

   There are 4 levels for education and 6 for age. It is a good idea to combine some groups for your analysis, since when you look for interaction the numbers within strata may be very small (you could check first whether this is likely to be a problem with the ‘tab’ command).

   To recode education into 2 levels:
   ```
   gen ed2 = ed
   recode ed2 2/4 = 2
   ```

   To recode age into 3 levels:
   ```
   gen age2 = age1
   recode age2 3=2 4/5=3 6=4
   ```

   First, examine the effects of education by age using `mhodds`:

   `mhodds case ed2, by(age2) c(2,1)`

   Maximum likelihood estimate of the odds ratio
   Comparing ed2==2 vs ed2==1
   by age2

<table>
<thead>
<tr>
<th>age2</th>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.519231</td>
<td>1.02</td>
<td>0.3132</td>
<td>0.142000  1.898596</td>
</tr>
<tr>
<td>2</td>
<td>1.952381</td>
<td>4.10</td>
<td>0.0430</td>
<td>1.009050  3.777605</td>
</tr>
<tr>
<td>3</td>
<td>3.458647</td>
<td>16.76</td>
<td>0.0000</td>
<td>1.836539  6.513465</td>
</tr>
<tr>
<td>4</td>
<td>2.847222</td>
<td>3.05</td>
<td>0.0808</td>
<td>0.833046  9.733360</td>
</tr>
</tbody>
</table>

   Mantel-Haenszel estimate controlling for age2
<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.289748</td>
<td>17.94</td>
<td>0.0000</td>
<td>1.543506  3.396777</td>
</tr>
</tbody>
</table>

   Test of homogeneity of ORs (approx): chi2(3) = 8.03
   Pr>chi2 = 0.0455
The test for homogeneity suggests that there may be some interaction between education and age in their effects on HIV infection.

Now, use logistic regression to test for interaction. First produce a model including interaction between ed2 and age2.

```
xi: logistic case i.ed2*i.age2
```

```
i.ed2             _Ied2_0-1           (naturally coded; _Ied2_0 omitted)
i.age2            _Iage2_1-4          (naturally coded; _Iage2_1 omitted)
i.ed2*i.age2      _Ied2Xage_#_#      (coded as above)
```

Logistic regression

```
Number of obs   =        763
LR chi2(7)      =      52.79
Prob > chi2     =     0.0000
Log likelihood = -400.73216                         Pseudo R2       =     0.0618
------------------------------------------------------------------------------
case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
    _Ied2_2 |   .5192308   .3402793  -1.00   0.317      .143722    1.875848
    _Iage2_2 |   1.3125   .8278301   0.43   0.666     .3812615    4.518307
    _Iage2_3 |   .7434783   .4503311  -0.49   0.625     .2268236    2.436959
    _Iage2_4 |   .6585366   .4170591  -0.66   0.510     .1903285    2.278537
   _Ied2Xage_2 |   3.760141   2.765401   1.80   0.072     .8895837    15.89357
   _Ied2Xage_3 |   6.661097   4.833445   2.61   0.009     1.60655    27.61831
   _Ied2Xage_4 |   5.483539   4.930153   1.89   0.058     .9413708    31.94193
------------------------------------------------------------------------------
```

The interaction terms are shown in the last 3 rows of the output. The parameter estimate for the interaction between age-group 3 and education is most significant, $P=0.009$.

To test for the overall significance of the interaction you should use a likelihood ratio test.

```
est store A
xi: logistic case i.ed2 i.age2
est store B
lrtest A B
```

```
likelihood-ratio test     LR chi2(3) =   6.43
(Assumption: B nested in A)     Prob > chi2 =   0.0923
```

There is weak evidence of interaction. You may decide to present a summary M-H or stratified estimates. However, generally for this level of significance a summary measure would be appropriate.
2. **Stratum-specific odds ratio estimates**

Assuming there is interaction, you should present stratum-specific odds ratio estimates. There are a number of ways you can do this. For the estimates of the effect of education on HIV infection below a logistic model can be produced for each age stratum:

**xi: logistic case i.ed2 if age2==1**

```
i.ed2  _Ied2_1-2  (naturally coded; Ied2_1 omitted)
Logistic regression             Number of obs =      109
LR chi2(1) =                 0.94
Prob > chi2 = 0.3331
Log likelihood = -39.366762   Pseudo R2 = 0.0118
------------------------------------------------------------------------------
case |   Odds Ratio   Std. Err.     z    P>|z|  [95% Conf. Interval]  
---------+---------------------------------------------------------------
   _Ied2_2 |  .5192308   .3402791  -0.317   0.751   .1437221   1.875846
------------------------------------------------------------------------------
```

**xi: logistic case i.ed2 if age2==2**

```
i.ed2  _Ied2_1-2  (naturally coded; Ied2_1 omitted)
Logistic regression             Number of obs =       288
LR chi2(1) =                 4.32
Prob > chi2 = 0.0376
Log likelihood = -181.1547    Pseudo R2 = 0.0118
------------------------------------------------------------------------------
case |   Odds Ratio   Std. Err.     z    P>|z|  [95% Conf. Interval]  
---------+---------------------------------------------------------------
   _Ied2_2 |  1.952381   .6516434   2.985   0.003   1.014986   3.755516
------------------------------------------------------------------------------
```

**xi: logistic case i.ed2 if age2==3**

```
i.ed2  _Ied2_1-2  (naturally coded; Ied2_1 omitted)
Logistic regression             Number of obs =       255
LR chi2(1) =                 17.11
Prob > chi2 = 0.0000
Log likelihood = -134.01229    Pseudo R2 = 0.0600
------------------------------------------------------------------------------
case |   Odds Ratio   Std. Err.     z    P>|z|  [95% Conf. Interval]  
---------+---------------------------------------------------------------
   _Ied2_2 |  3.458647   1.077418   3.198   0.001   1.878213   6.368947
------------------------------------------------------------------------------
```
8.4 Practical 8: Solutions

**xi: logistic case i.ed2 if age2==4**

```
i.ed2  _Ied2_1-2  (naturally coded; _Ied2_1 omitted)
Logistic regression                     Number of obs =   111
LR chi2(1) =       2.65
Prob > chi2 =    0.1035
Log likelihood = -46.198409             Pseudo R2 =   0.0279

case  | Odds Ratio  Std. Err.     z  P>|z|     [95% Conf. Interval]
----+---------------------------------------------------------------
  _Ied2_2 |    2.847222  1.752509  1.70  0.089    .8521062    9.51369
```

Alternatively, a single model can be fitted with an interaction term:

**xi: logistic case i.ed2*i.age2**

```
i.ed2  _Ied2_0-1  (naturally coded; _Ied2_0 omitted)
i.age2  _Iage2_1-4  (naturally coded; _Iage2_1 omitted)
i.ed2*i.age2  _Ied2Xage_#_#  (coded as above)
Logistic regression                     Number of obs =   763
LR chi2(7) =      52.79
Prob > chi2 =     0.0000
Log likelihood = -400.73216             Pseudo R2 =   0.0618

case  | Odds Ratio  Std. Err.     z  P>|z|     [95% Conf. Interval]
----+---------------------------------------------------------------
   _Ied2_2 |     0.519231   .3402793 -1.00  0.317   .1437222    1.875848
   _Iage2_2 |      1.3125   .8278301  0.43  0.666   .3812615   4.518307
   _Iage2_3 |      0.743479   .4503111 -0.49  0.625   .2268235   2.436959
   _Iage2_4 |      0.658536   .4705911 -0.66  0.510   .1903285   2.278537
   _Ied2Xage_2 |      3.760141  2.7654014  1.80  0.072   .8895837   15.89357
   _Ied2Xage_3 |      6.661097  4.8334445  2.61  0.009   1.606550   27.61831
   _Ied2Xage_4 |      5.483539  4.9301534  1.89  0.064   .9413708   31.94193
```

And Stata’s **lincom** command can be used to compute stratum-specific odds ratios with 95% confidence intervals:

**lincom _Ied2_2 + _Ied2Xage_2_2**

```
case  | Odds Ratio  Std. Err.     z  P>|z|     [95% Conf. Interval]
----+---------------------------------------------------------------
  (1) |      1.952381   .6516434  2.00  0.045    1.014986    3.75551
```

**lincom _Ied2_2 + _Ied2Xage_2_3**

```
case  | Odds Ratio  Std. Err.     z  P>|z|     [95% Conf. Interval]
----+---------------------------------------------------------------
  (1) |      3.458647   1.077422  3.98  0.000    1.878216    6.368954
```

**lincom _Ied2_2 + _Ied2Xage_2_4**

```
case  | Odds Ratio  Std. Err.     z  P>|z|     [95% Conf. Interval]
----+---------------------------------------------------------------
  (1) |      2.847222  1.7525155  1.70  0.089    .8521027    9.51373
```
This method is preferable because each estimate is based on the full dataset, although in this example the confidence intervals are similar to those obtained using a separate model for each stratum.

When presenting the results from this model, a table can be produced with the following effects reported (this is a demonstration table, and the results could be presented in other ways):

<table>
<thead>
<tr>
<th>Stratum-specific effects of education in each agegroup:</th>
<th>Adjusted OR*</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age group 1 (baseline agegroup)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education 0 (baseline)</td>
<td>1.00</td>
<td>0.14, 1.88</td>
<td>0.32</td>
</tr>
<tr>
<td>Education level 1</td>
<td>0.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age group 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education 0 (baseline)</td>
<td>1.00</td>
<td>1.01, 3.76</td>
<td>0.05</td>
</tr>
<tr>
<td>Education level 1</td>
<td>1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age group 3</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Education 0 (baseline)</td>
<td>1.00</td>
<td>1.88, 6.37</td>
<td></td>
</tr>
<tr>
<td>Education level 1</td>
<td>3.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age group 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education 0 (baseline)</td>
<td>1.00</td>
<td>0.85, 9.51</td>
<td>0.09</td>
</tr>
<tr>
<td>Education level 1</td>
<td>2.85</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Main effect of age group, in the baseline level of education:</th>
<th>Adjusted OR*</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age group 1 (baseline)</td>
<td>1.00</td>
<td>0.89, 15.89</td>
<td>0.07</td>
</tr>
<tr>
<td>Age group 2</td>
<td>3.76</td>
<td>1.61, 27.62</td>
<td>0.01</td>
</tr>
<tr>
<td>Age group 3</td>
<td>6.66</td>
<td>0.94, 31.94</td>
<td>0.06</td>
</tr>
<tr>
<td>Age group 4</td>
<td>5.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>If other variables are included in final model they can be included in the table:</th>
<th>Adjusted OR*</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genital ulcer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not present (baseline)</td>
<td>1.00</td>
<td>etc</td>
<td>etc</td>
</tr>
<tr>
<td>Present</td>
<td>aa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Moslem (baseline)</td>
<td>1.00</td>
<td>etc</td>
<td>etc</td>
</tr>
<tr>
<td>2 Catholic</td>
<td>bb</td>
<td></td>
<td>etc</td>
</tr>
<tr>
<td>3 Protestant</td>
<td>cc</td>
<td></td>
<td>etc</td>
</tr>
<tr>
<td>4 Other</td>
<td>dd</td>
<td></td>
<td>etc</td>
</tr>
</tbody>
</table>

* Adjusted for all variables shown in table.

**Summary**

In this exercise you have

- Used the **logit** command in Stata to perform a logistic regression analysis of the relationship between HIV infection, age and education, including an interaction term between education and age.
- Performed the likelihood ratio test in Stata using the **lrtest** command to determine whether or not the interaction term between age and education should be included in the model for HIV infection
- Obtained stratum-specific odds ratio estimates in the analysis of the effect of age and education on the odds of HIV infection, assuming the interaction term is necessary.
Practical 9

Solutions

You should first open a log file to save your results.

\texttt{e.g. \texttt{log using smeex9.log}}

Read in the mwanza dataset.

\texttt{use mwanza, clear}

1. \textit{HIV infection and the number of infections in the past year}

\texttt{tab case inj}

\begin{verbatim}
<table>
<thead>
<tr>
<th>Case/control</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>278</td>
<td>53</td>
<td>124</td>
<td>83</td>
<td>35</td>
<td>1</td>
<td>574</td>
</tr>
<tr>
<td>1</td>
<td>68</td>
<td>15</td>
<td>41</td>
<td>41</td>
<td>23</td>
<td>1</td>
<td>189</td>
</tr>
<tr>
<td>Total</td>
<td>346</td>
<td>68</td>
<td>165</td>
<td>124</td>
<td>58</td>
<td>2</td>
<td>763</td>
</tr>
</tbody>
</table>
\end{verbatim}

The two missing values for \texttt{inj} should be recoded to missing:

\texttt{recode inj 9=.}

\texttt{tab case inj, col}

\begin{verbatim}
<table>
<thead>
<tr>
<th>Case/control</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80.35</td>
<td>77.94</td>
<td>75.15</td>
<td>66.94</td>
<td>60.34</td>
<td>75.30</td>
</tr>
<tr>
<td>1</td>
<td>19.65</td>
<td>22.06</td>
<td>24.85</td>
<td>33.06</td>
<td>39.66</td>
<td>24.70</td>
</tr>
<tr>
<td>Total</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
\end{verbatim}

The proportion of cases increases with increasing numbers of injections.
2. **Odds of HIV infection for each category of injections**

```
tabodds case inj
```

<table>
<thead>
<tr>
<th>inj</th>
<th>cases</th>
<th>controls</th>
<th>odds</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>278</td>
<td>0.24460</td>
<td>0.18763  0.31888</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>53</td>
<td>0.28302</td>
<td>0.15954  0.50207</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>124</td>
<td>0.33065</td>
<td>0.23228  0.47066</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>83</td>
<td>0.49398</td>
<td>0.33980  0.71811</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>35</td>
<td>0.65714</td>
<td>0.38831  1.11209</td>
</tr>
</tbody>
</table>

Test of homogeneity (equal odds): \(\chi^2(4) = 16.61\)

\(\text{Pr}>\chi^2 = 0.0023\)

Score test for trend of odds: \(\chi^2(1) = 15.36\)

\(\text{Pr}>\chi^2 = 0.0001\)

The odds of HIV infection increase with increasing number of injections.

3. **Logistic model assuming a linear effect**

```
logistic case inj
```

Logistic regression

```
Logistic regression                                     Number of obs =    761
chi2(1)       =  15.11
Prob > chi2   = 0.0001
Log Likelihood = -417.88993                             Pseudo R2     = 0.0178
```

```
case | Odds Ratio | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
-----|------------|-----------|-------|------|----------------------|
inj  | 1.263075   | .0759265  | 3.885 | 0.000| 1.122694  1.421009  |
```

This model includes the group with zero injections. Because this group can differ from the rest, and to make sure you are testing for a trend with increasing exposure, produce a model excluding the group with no injections in the past year.

```
logistic case inj if inj!=1
```

Logistic regression

```
Logistic regression                                     Number of obs =    415
LRchi2(1)       =   6.84
Prob > chi2   = 0.0089
Log Likelihood = -246.15766                             Pseudo R2     = 0.0137
```

```
case | Odds Ratio | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
-----|------------|-----------|-------|------|----------------------|
inj  | 1.36138    | .1617354  | 2.597 | 0.009| 1.078586  1.718321  |
```

The common OR is 1.36.

For every increase in the level of number of injections the odds of HIV infection increases by 1.36. After excluding the zero group, there remains a highly statistically significant linear trend for increasing odds of HIV infection with increasing number of injections in the past year.
4. **Test for departure from a linear trend**

To test for departure from a linear trend in the effect of number of injections you use a likelihood ratio test to compare the model with a separate effect for each level of inj with the model that assumes a common OR.

\[ \text{xi: logistic case inj if inj!=1} \]

```
Logistic regression                                   Number of obs   =        415
LR chi2(3)      =       7.08
Prob > chi2     =     0.0693
Log likelihood = -246.03677                       Pseudo R2       =     0.0142
------------------------------------------------------------------
case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
   _Iinj_3 |    1.16828   .4012991     0.45   0.651     .5958853    2.290503
   _Iinj_4 |   1.745382   .6095628     1.59   0.111     .8802607    3.460744
   _Iinj_5 |   2.321905   .9217239     2.12   0.034     1.066458    5.055276
------------------------------------------------------------------
est store A
logistic case inj if inj!=1

Logistic regression                                   Number of obs   =        415
LR chi2(1)      =       6.84
Prob > chi2     =     0.0089
Log likelihood = -246.15766                       Pseudo R2       =     0.0137
------------------------------------------------------------------
case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
      inj |    1.36138   .1617354     2.60   0.009     1.078586    1.718321
------------------------------------------------------------------
est store B
lrtest A B
likelihood-ratio test (Assumption: B nested in A)
LR chi2(2) =    0.24
Prob > chi2 =    0.8861
```

P=0.89, so there is no evidence against the null hypothesis of a linear trend. In other words, there is no evidence for a departure from a linear trend. The logistic model assuming a linear trend is, therefore, appropriate, although the z statistic is somewhat reduced after excluding the zero group.

Note: This result is very difficult to interpret. One reason for injections is because the subject had an HIV-related illness. It does not show that injection is a major mode of transmission of HIV in Mwanza!

**Summary**

In this exercise you have

- Used the logit command for HIV infection that assumes a linear effect of number injections.
- Used the likelihood ratio test command lrtest to test for a departure from a linear trend in the effect of number of injections.
9.4 Practical 9: Solutions
Section 4: Solutions

Practical 10

Solutions

Getting Started

You should first open a log file to save your results.

\texttt{e.g. log using asmex2.log}

1. \textbf{What is the effect of birth weight (bwtgp) on infant mortality due to diarrhoea?}

\texttt{use diabraz}
\texttt{help diabraz}
\texttt{match case bwtgp pair}

\begin{center}
\begin{tabular}{lrrr}
& 0 bwtgp & 1 bwtgp & Total \\
\hline
1 & 31 & 18 & 49 \\
2 & 25 & 12 & 37 \\
\hline
Total & 56 & 30 & 86 \\
\end{tabular}
\end{center}

\texttt{mhoods case bwtgp pair, c(2,1)}

\begin{center}
\begin{tabular}{rccc}
Odds ratio & chi2(1) & P>chi2 & [95\% Conf. Interval] \\
\hline
1.388889 & 1.14 & 0.2858 & 0.757781 - 2.545608 \\
\hline
\end{tabular}
\end{center}

There is no evidence for an effect of birthweight on mortality from diarrhoea.
2. Is there any evidence of interaction between birth weight and age? (Use age in 2 groups 0-2 month, 3+ months.)

This is the new variable created during Practical 2. We use two groups rather than more to improve the chance of showing an interaction if one exists, since the number of strata in the analysis is small.

```
match case bwtgp pair if agenew==1
<table>
<thead>
<tr>
<th>0 bwtgp</th>
<th>1 bwtgp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
</tr>
</tbody>
</table>

match case bwtgp pair if agenew==2
<table>
<thead>
<tr>
<th>0 bwtgp</th>
<th>1 bwtgp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
</tr>
</tbody>
</table>
```

```
mhodds case bwtgp pair if agenew==1, c(2,1)
Mantel-Haenszel estimate controlling for pair

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.166667</td>
<td>0.08</td>
<td>0.7815</td>
<td>0.392087 3.471456</td>
</tr>
</tbody>
</table>

mhodds case bwtgp pair if agenew==2, c(2,1)
Mantel-Haenszel estimate controlling for pair

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.500000</td>
<td>1.20</td>
<td>0.2733</td>
<td>0.722549 3.113977</td>
</tr>
</tbody>
</table>
```

Use the numbers from the above tables to form a table of discordant pairs.

```
<table>
<thead>
<tr>
<th>Birth weight</th>
<th>Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-2 months</td>
</tr>
<tr>
<td>Case + control -</td>
<td>7</td>
</tr>
<tr>
<td>Case – control +</td>
<td>6</td>
</tr>
</tbody>
</table>
```
tabi 7 18 \[6 12, \text{chi}^{2}\]

<table>
<thead>
<tr>
<th>row</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>30</td>
<td>43</td>
</tr>
</tbody>
</table>

Pearson chi2(1) = 0.1411 Pr = 0.707

Again, there is no evidence of interaction between birthweight and age group. That is, there is no support for the hypothesis that the effect of low birthweight on death from diarrhoea is different in infants of different ages.

3. **What is the effect of birth weight in this dataset?**

Using `diabraz2`: (Note: the `match` command will not work with more than one control per case but we can still obtain the matched odd ratio.)

```
mhodds case bwtgp set, c(2,1)
```

Mantel-Haenszel estimate controlling for set

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.637931</td>
<td>6.11</td>
<td>0.0134</td>
<td>1.103241 2.431760</td>
</tr>
</tbody>
</table>

This shows that the odds of mortality is 1.64 times higher in infants with "low" birthweight (bwtg=2 i.e. birthweight <3kg) compared with infants with "normal" birthweight (≥3kg).

4. **Is there any evidence of interaction with age as defined above?**

```
mhodds case bwtgp set if agenew==1, c(2,1)
```

Mantel-Haenszel estimate controlling for set

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.222222</td>
<td>0.14</td>
<td>0.7103</td>
<td>0.423112 3.530575</td>
</tr>
</tbody>
</table>

```
mhodds case bwtgp set if agenew==2, c(2,1)
```

Mantel-Haenszel estimate controlling for set

<table>
<thead>
<tr>
<th>Odds ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.608696</td>
<td>3.37</td>
<td>0.0665</td>
<td>0.963450 2.686077</td>
</tr>
</tbody>
</table>

You cannot do a test for interaction with more than 1 control per case using classical methods. Looking at the odds ratio in each stratum does not show a large difference between them, and the confidence intervals overlap.