London School
of Hygiene &
Tropical Medicine

STATA GLOSSARY
and
STATA TUTORIAL

STATISTICAL METHODS IN EPIDEMIOLOGY
and
ADVANCED STATISTICAL METHODS IN EPIDEMIOLOGY

2004

As at: 09/01/2004

Statistical Methods in Epidemiology (2402) 2003-04

Practical Facilitators & Rooms

Date	Time		Practical Number & Title	Group 1		Group 2		Group 4		Group 5		Group 6		Group 7		Group 3		Group 8		Group 9	
				Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator	Room	Facilitator
14/01/2004	3.30pm	1	Use of STATA for calculation of measures of effect (Computer Practical)	99/1	SC*	M/LG7	VM*	99/2	BS*	LG3	MP*	LG2	EW	LG2	CH*	LG31	AH	LG31	CC*		
15/01/2004	11.00am	2	Crude & Stratified Rates (Computer Practical)	99/1	LS*	M/LG7	SC*	99/2	TE*	LG3	MP*	LG2	CC*	LG2	CT	LG31	BS*	LG31	AH		
	3.30pm	3	Survival Analysis (Computer Practical)	99/1	LS*	M/LG7	CH*	99/2	TE*	LG3	MP*	LG2	AH	LG2	BS*	LG31	SC*	LG31	CT		
16/01/2004	11.00am	4	Case-Control Studies (Non-Computer Practical)	152	BR*	153	CH*	509	EB*	633	SC*	B04	SF*	B13	TE*	328	STh*	B14	CT*	Council Room	PH*
	2.00pm	5	Data Checking/Editing & Univariate Analysis (Computer Practical - extd. To 5pm)	365	CC*	99/1	PH*	99/2	SC*	LG3	TE*	LG2	AH	LG2	CH*	LG31	SF*	LG31	BR		
21/01/2004	2.00pm	OR	Optional Review Session	Liam Smeeth																	
22/01/2004	11.00am	6	Analysis of Unmatched Case Control Studies (Computer Practical)	365	CS*	M/LG7	NA*	99/2	SF*	LG3	BR*	LG2	KF*	LG2	AH	LG31	LS	LG31	SC*		
	3.30pm	7	Private Study	N/A																	
23/01/2004	11.00am	8	Likelihood (Computer Practical)	99/1	ST*	M/LG7	SF*	99/2	PH*	LG3	BR*	LG2	SC*	LG2	CS	LG31	KF*	LG31	AH		
	3.30pm	9	Use of STATA for simple logistic regression analyses (Computer Practical)	99/1	JT*	M/LG7	EB*	99/2	NA*	LG3	ST*	LG2	CS	LG2	SC*	LG31	PH*	LG31	EW		
28/01/2004	2.00pm	OR	Optional Review Session	Polly Hardy																	
29/01/2004	11.00am	10	Logistics Regression 2:Models with more than one variable (Computer Practical)	365	EB*	99/1	SC*	99/2	PH*	LG3	LS*	LG2	ST*	LG2	EW	LG31	JT*	LG31	AS		
	3.30pm	11	Logistics Regression 3 (Computer Practical)	99/1	SC*	M/LG7	BR*	99/2	JT*	LG3	ST*	LG2	AS	LG2	PH*	LG31	LS*	LG31	AH		
30/01/2004	11.00am	12	Logistics Regression 4 (Computer Practical)	99/1	JT*	M/LG7	ST*	99/2	BR*	LG3	PH*	LG2	SC*	LG2	AS	LG31	EB*	LG31	EW		
	3.30pm	13	Private Study	N/A																	
04/02/2004	2.00pm	OR	Optional Review Session	Simon Cousens																	
05/02/2004	11.00am	14	Analyses of Matched Case-Control Studies (Computer Practical)	99/1	VM*	M/LG7	NA*	99/2	JT?*	LG3	SC*	LG2	AS	LG2	MS*	LG31	AH	LG31	LS*		
	3.30pm	15	Strategies of Analysis - (Computer Practical)	365	MS*	M/LG7	LS*	99/2	NA*	LG3	SC*	LG2	CT	LG2	VM*	LG31	EB*	LG31	AH		
06/02/2004	9.30am	16	Presentation of Statistical Results: Group Work (Non Computer Practical)	251	EB*	153	PH*	253	CS*	254	BR*	255	VM*	405	NA*	252	SC*	406	CT*	407	SF*
11/02/2004																					
12/02/2004			PRIVATE STUDY																		
13/02/2004																					

Computer Practicals - Student Numbers per Room

Room 365	22
Room 99/1	When used by Group 1 - 22, When used by Group 2 - 25
Room 99/2	25
Room LG3 (originally B202)	20
Room LG2 (originally B203)	38
Room LG31 (originally B9)	38
Room M/LG7	25

* = Practical Leader

Please note: Non-Computer Practical rooms are all based at Birkbeck College, Malet Street. (The first digit of the room number corresponds to which floor the room is situated on in the building)

Statistical Methods in Epidemiology (2402) 2003-04

Practical Facilitators & Rooms

Facilitators		Facilitating Hours
Alexander, Neal	NA	9
Breeze, Elizabeth	EB	10.5
Cook, Claire	CC	6
Cousens, Simon	SC	24
De Stavola, Bianca	BS	4.5
Edwards, Tansy	TE	7.5
Fielding, Katherine	KF	3
Floyd, Sian	SF	10.5
Haghdoost, Ali-Akbar	AH	15
Hardy, Polly	PH	15
Higgins, Craig	CH	7.5
McCormack, Valerie	VM	7.5
Payne, Mary	MP	4.5
Rachet, Bernard	BR	13.5
Schoemaker, Minouk	MS	3
Shah, Anjali	AS	6
Sismanidis, Charalambos	CS	7.5
Smeeth, Liam	LS	10.5
Tam, Clarence	CT	9
Thomas, Sara	STh	1.5
Todd, Jim	JT	7.5
Tomkins, Susannah	ST	7.5
Williamson, Elizabeth	EW	6

Please Note: Extended Session (16th December) counts as 3 hours / 2 practicals.
Please Note: Double Session (6th February) counts as 3 hours / 2 practicals.

Statistical Methods in Epidemiology (2402)

Student List and Practical Groups

Surname	Forename	MSc	Non-Computing Practical Group	Computing Practical Group
ABDULLAH MUDA	NORAIDATULAKMA	MS	4	1
AJANGA	ANTONY	CID	2	2
ALBANESE	EMILIANO	PHN	7	2
ALBERTI	KATE	EPI	7	1
ALFRED	TAMUNO	MS	5	5
ARCHER	CAROLYN	EPI	8	6
ASHE	SEAN	VET	7	7
ATSBEHA	TESMERELNA	PHDC	1	8
BAILEY	LESLEY	MS	6	1
BANSI	LOVELEEN	MS	7	2
BARD	ELLIE	EPI	9	2
BARDACH	ARIEL	EPI	1	4
BATCHELOR	NICOLA	CID	3	5
BENNETT	RACHEL	MS	8	6
BERTRAN SERRA	MARIA	VET	8	7
BHADHAL	HARDIP	PHN	8	8
BINDRA	RENU	PH	6	1
BISHOP	LOUISE	EPI	2	8
BLACK	KIRSTEN	RES-PHP	N/A	N/A
BOHLIUS	JULIA	PH	7	1
BOSTOEN	KRISTOF	RES-ITD	N/A	N/A
BOZICEVIC	IVANA	RES-PHP	N/A	N/A
BRADSHAW	KATE	MS	9	6
BRISTOW	KIRSTY	EPI	3	7
BROWN	ANDREW	MS	1	8
BUITRAGO JARAMILLO	JULIANA	EPI	4	3
BUKASA	ANTOANETA	MS	2	2
CAMPBELL	LUCY	EPI	5	2
CARDENAS SANCHEZ	ANGELA	EPI	6	4
CARPENTER	ANNA	PHN	9	5
CARRINGTON	JOANNA	MS	3	6
CASTANON WILLIAMS	ALEXANDRA	EPI	7	7
CESARONI	GIULIA	EPI	8	8
CHIBWESHA	CARLA	EPI	9	1
CHIMED	AL	OCC-EPH	3	2
CHISSTER	IRINA	MS	4	2
CLARK	EMMA	EPI	1	2
CLARKE	JUDITH	MS	5	5
ARIBABA	OLUFISAYO	CEH	2	6
COSTA	HELIO	OCC-EPH	4	7
COSTARD	SOLENNE	VET	9	8
CROWE	SAMUEL	PH	8	1
CUMMINS	STEVEN	EPI	2	2
D'AGUIAR	STEPHEN	MS	6	2
DAMOAH	KWAKU	MS	1	4
DATTA	PREETI	HS/SR	2	5
DAVIES	ANNA	RES-EPH	N/A	N/A
DE MENEZES	CLAIRE	PHDC	3	6
DE ROZA	DY-JUAN	EPI	3	7
DELLICOUR	STEPHANIE	EPI	4	8
DIESEL	GILLIAN	VET	1	1
DIMITRIOVA	BOIKA	RES-PHP	N/A	N/A

Student List and Practical Groups

Surname	Forename	MSc	Non-Computing Practical Group	Computing Practical Group
DOOBAREE	INDRARAJ	EPI	5	2
EAPEN	KOSHY	EPI	6	4
ENYEGUE OYE	JOSEPH	CEH	1	5
ESCRIBANO FERRER	BLANCA	EPI	7	6
FARLEY-HESS	CLAUDIA	PH	9	7
FERRARO	ALEXANDRE	EPI	8	8
FITZGERALD	MOLLY	RES-EPH	N/A	N/A
FORDE	JOSH	DH	1	2
GEORGE	JULIE	PH	1	2
GIBSON	JACK	EPI	9	4
GOPAL	RACQUEL	VET	2	5
GORGOS	LINDA	CID	4	6
GUPTA	AJAY	EPI	1	7
HAMILTON	KEITH	CID	5	8
HANLON	CHARLOTTE	EPI	2	1
HARDOON	SARAH	MS	7	8
HEAD	SARAH	PH	2	8
HEISS	LORI	RES-ITD	N/A	N/A
HASSAAN	FOUAD	RES-EPH	N/A	N/A
MAK	TIPPI	PH	5	8
HAWKESWORTH	SOPHIE	PHN	6	7
HERNANDEZ-SUAREZ	GUSTAVO	EPI	3	6
HILDEBRANDT	RUTH	PHDC	4	7
HOSSAIN	MAZEDA	RSHR	2	8
HOUSTON	CHRISTOPHER	VET	3	4
INMACULADA ASENSIO	JOSEFA	VET	4	2
ANDREW	MELISSA	PH	4	2
KALAMPOLI	VASILIKI	MS	8	4
KALER	JASMEET	VET	5	5
KAMPHUIS	MONIQUE	PHDC	5	6
KAPOSVARI	CSILLA	EPI	5	7
KAY	CHRISTINA	RSHR	1	8
KAYE	SAMANTHA	EPI	6	4
KELLY	LESLIE	RES-PHP	N/A	N/A
KHANMAYO	AQSAA	MS	9	2
KITCHEN	LISA	EPI	7	4
KODAMA	TOMOKO	PH	3	5
KOMOTO	SHIGEKAZU	EEP	6	6
KOOLE	OLIVIER	PHDC	6	7
KOSBAEVA	ALIA	PHDC	7	8
KU	YOUNG	TMIH	6	4
LAHUERTA-MARIN	ANGELA	VET	6	2
LEE	VIVIAN	DH	2	2
LEMMA JIMA	MISIKIR	EPI	8	4
LEMME	FRANCESCA	MS	2	5
LEVY	GUS	MS	3	6
LEWANDOWSKI	ERIC	EPI	9	7
LIM	ERIC	MS	4	8
LOUIE	KARLY	EPI	1	1
MANN	ANDREA	EPI	2	2
MAX	VANESSA	VET	7	2
MCCABE	PAUL	PH	4	4

Student List and Practical Groups

Surname	Forename	MSc	Non-Computing Practical Group	Computing Practical Group
MCCAMBRIDGE	JAMES	EPI	3	5
MCDEVITT	JOSEPH	MS	5	6
MCGUIRE	MEGAN	EPI	4	7
MICAH	FRANK	EPI	5	8
MILLS	NICOLA	EPI	6	1
MINH	PHAN	VET	8	7
MONSEES	GENEVIEVE	MS	6	2
MOONEN	BRUNO	EPI	7	4
MORGAN	CRAIG	HS/SR	3	5
MORI	RINTARO	EPI	8	6
MUNGUAMBE	KHATIA	RES-ITD	N/A	N/A
MUNRO	HELEN	EPI	9	8
MURIUKI	DAVID	PHDC	8	1
MUSTON	DOMINIC	MS	7	2
MWAUNGULU	FRANK	EPI	1	7
NAPP	SEBASTIAN	VET	9	6
NASCIMENTO	MARIA	RES-ITD	N/A	N/A
NEWEY	CLAIRE	EPI	2	6
NISHIKIORI	NOBUYUKI	EPI	3	7
OBARA	HIROMI	EPI	4	8
ODEK	WILLIS	DH	3	1
OKELL	LUCY	EPI	5	2
OSMAN SAEED	MAHA	RES-ITD	N/A	N/A
OYEE	JAMES	MS	8	7
OZGEDIZ	DORUK	PHDC	9	5
PANDEY	POOJA	PHDC	1	6
PAPADPOULOU	CHRISTINA	VET	1	7
PARKER	JULIA	PHDC	7	6
PATEL	BELA	EPI	6	3
PEER	ANESSA	EPI	7	1
PEREL	PABLO	EPI	8	4
PEREZ-ACHIAGA	NATALIA	EPI	9	5
PHILLIPS	CAROLINE	EPI	1	3
PHILLIPS	PAMELA	EPI	2	1
PICADO	ALBERT	VET	2	4
PITHUA	PATRICK	VET	3	5
RANGAKA	MOLEBOGENG	EPI	3	3
RIGGS-NAGY	JAMIE	CID	6	1
RIGHARTS	ALIDA	EPI	4	4
RITCHIE	SARA	PHDC	2	3
RIVERA-MARQUEZ	ALBERTO	RES-EPH	N/A	N/A
ROBERTS	ALWYN	VET	4	4
ROKICKI	JESSE	RSHR	3	3
ROTTER	RUTH	CID	7	1
RUBINO	ANNALISA	EPI	5	4
RUNDI	CHRISTINA	RES-ITD	N/A	N/A
SAHOO	DIPTIKANTI	PH	5	1
SALEHEEN	SARAH	CID	8	4
SAN MIGUEL	BEATRIZ	STI	5	3
SEARLE	KENDALL	EPI	6	1
SESTAK	IVANA	OCC-PHP	5	3
SEWARD	NADINE	EPI	7	3

Student List and Practical Groups

Surname	Forename	MSc	Non-Computing Practical Group	Computing Practical Group
SEYLER	THOMAS	CID	9	1
SHEIKH	SUHAIL	PH	6	4
SHOWELL	DANIEL	PHDC	3	3
SIDDIQUI	MAHVEEN	PHDC	4	1
SINGH	GITA	RSHR	4	4
SIZAIRE	VINCIANE	EPI	8	3
SMITH	JENNIFER	EPI	9	4
SMITH	MATTHEW	PH	7	3
SOLEMAN	NADIA	PHDC	5	5
SPRINGER	KAREN	EPI	1	3
SQUIRE	NICOLA	EPI	2	4
STAEDKE	SARAH	RES-ITD	N/A	N/A
SWEENEY DOW	ELIZABETH	MS	9	3
THOMPSON	PAULA	EPI	3	5
TOL	BUNKEA	EPI	4	3
TOMIO	JUN	PH	8	5
TSUZUKI	ATARU	EPI	5	3
TURNER	RACHEL	PH	9	5
VYSE	ANDREW	EPI	6	3
WALTERS	PAUL	EPI	7	5
WATTS	CHARLOTTE	STAFF	N/A	N/A
WELLS	CHRISTINE	MS	1	4
WILLIS	RUTH	EPI	8	3
WOODCOCK	JAMES	RES-EPH	N/A	N/A
WORTON	SARAH	MS	2	4
XAVIER VALLES CASANOVA	FRANCESCA	EPI	9	1
YEATMAN	SARA	DH	5	2
YIP	JENNIFER	EPI	1	3
YOUNG	CYNTHIA	PHDC	6	6

ROOM ALLOCATIONS

There are nine practical groups for non-computer practicals and eight for computer practicals. Most groups include people from a variety of MScs. Please take the opportunity to share skills and experiences. All Research Students and Staff attending the course are required to complete the practicals at their own computers and therefore have not been assigned to practical groups.

The timetable indicates whether a practical takes place in computer rooms or non-computer rooms. The rooms for all groups practicals are listed on the 'Practical' Locations sheet in this folder.

We are unable to reserve pcs for your use during the assessment week except Thursday 12 February & Friday 13 February, when rooms 99/1, 99/2 and M/LG7 are booked 9.30-12.30.

Statistical Methods in Epidemiology - Practical Timetable 2003-04

Date	Time	Practical Number	Group 1 Room	Group 2 Room	Group 3 Room	Group 4 Room	Group 5 Room	Group 6 Room	Group 7 Room	Group 8 Room	Group 9 Room
14/01/2004	3.30pm	1	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
15/01/2004	11.00am	2	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	3	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
16/01/2004	11.00am	**4 (Non Computing)**	**152**	**153**	**328**	**509**	**633**	**B04**	**B13**	**B14**	**Council Room**
	2.00pm	5	365	99/1	LG31	99/2	LG3	LG2	LG2	LG31	N/A
21/01/2004	2.00pm	OR	GOLDSMITHS								
22/01/2004	11.00am	6	365	99/1	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	7	PRIVATE STUDY								
23/01/2004	11.00am	8	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	9	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
28/01/2004	2.00pm	OR	GOLDSMITHS								
29/01/2004	11.00am	10	365	99/1	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	11	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
30/01/2004	11.00am	12	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	13	PRIVATE STUDY								
04/02/2004	2.00pm	OR	GOLDSMITHS								
05/02/2004	11.00am	14	99/1	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
	3.30pm	15	365	M/LG7	LG31	99/2	LG3	LG2	LG2	LG31	N/A
06/02/2004	9.30am	**16 (Non Computing)**	**251**	**153**	**252**	**253**	**254**	**255**	**405**	**406**	**407**
11/02/2004											
12/02/2004		PRIVATE STUDY									
13/02/2004											

We are unable to reserve pcs for your use during the assessment week except Thursday 12 February & Friday 13 February, when rooms 99/1, 99/2 and M/LG7 are booked 9.30-12.30.

Please note: Computer Practical rooms are all located at LSHTM. Non-Computer Practical rooms are located at Birkbeck College, Malet Street. (Birkbeck Rooms - The first digit of the room number corresponds to which floor the room is situated on in the building)

Location of LSHTM Computer Rooms

365	Keppel Street, 3rd Floor
99/1 and 99/2	Based at 99 Gower Street
LG2	Keppel Street, Lower Ground Floor
LG3	Keppel Street, Lower Ground Floor
LG31	Keppel Street, Lower Ground Floor
M/LG7	Based in The Mews (Lower ground), Bedford Square

Statistical Methods in Epidemiology (2402)

Course Handbook 2004

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

STUDENT EVALUATION QUESTIONNAIRE FOR TEACHING UNITS (2003-2004)

This evaluation form is in two parts. Section 1 provides you with an opportunity to comment on individual sessions during the teaching unit. We encourage you to make criticisms and/or positive comments, together with any suggestions for change. Section 2 provides you with an opportunity to comment on aspects of the course in general.

Name of teaching unit: **STATSITICAL METHODS IN EPIDEMIOLOGY (2402)**

Name of your MSc:

SECTION 1

Please use the table below to give feedback on individual sessions. Aspects you might wish to comment on include the quality of teaching, the content of the session, facilities, tutor support etc.

Date	Title of Session	Type	Comments
14/01/03	Introduction to SME	Lecture & Practical	
14/01/03	Summary of Measures of Effect	Lecture & Practical	
15/01/03	Crude and Stratified Rates	Lecture & Practical	
15/01/03	Survival Analysis	Lecture & Practical	
16/01/03	Case Control Studies – Concepts	Lecture & Practical	
16/01/03	Data Checking / Editing and Univariate Analysis	Practical	
22/01/03	Analysis of Case-Control Studies	Lecture & Practical	
22/01/03	Likelihood Theory	Lecture & Practical	
23/01/03	Approximate Likelihoods	Lecture & Practical	
23/01/03	Logistic Regressions 1	Lecture & Practical	
29/01/03	Logistic Regression 2	Lecture & Practical	
29/01/03	Logistic Regression 3	Lecture & Practical	
30/01/03	Logistic Regression 4	Lecture & Practical	
30/01/03	Matched Case-Control Studies	Lecture & Practical	
05/02/03	Analysis of Matched Case-Control Studies	Lecture & Practical	
05/02/03	Strategies of Analysis	Lecture & Practical	
06/02/03	Presentation of Statistical Results	Lecture & Practical	
06/02/03	Introduction to Assessment	Lecture	
21/01/03 28/01/03 04/02/03	Optional Reviews	Other	

STUDENT EVALUATION QUESTIONNAIRE FOR TEACHING UNITS

Name of teaching unit: **STATSITICAL METHODS IN EPIDEMIOLOGY (2402)**

Name of your MSc :

SECTION 2: Please circle one of the numbers in each line

	Very good	Good	Neutral	Bad	Very bad	Not applicable
1. Quality of course notes and reading material *Comments*	5	4	3	2	1	0
2. Quality of lectures *Comments*	5	4	3	2	1	0
3. Quality of practicals *Comments*	5	4	3	2	1	0
4. Content of course *Comments*	5	4	3	2	1	0
5. Fulfilment of course objectives *Comments*	5	4	3	2	1	0
6. Division of time between activities, including private study time *Comments*	5	4	3	2	1	0
7. Your overall opinion of the course	5	4	3	2	1	0

Overall, what were the good things about this teaching unit?:

What could be improved?:

Further pages may be added, if needed. Thank you for completing this questionnaire.

Please return this questionnaire to: Kirsty Ransome (Room 107, Keppel Street)

STATISTICAL METHODS IN EPIDEMIOLOGY 2004

COURSE HANDBOOK

CONTENTS

INTRODUCTION

The course on Statistical Methods in Epidemiology is designed to provide you with the skills needed to analyse and interpret cohort, case-control and cross-sectional studies by cross-tabulation, stratification and regression.

By the end of the unit students should be able to:
(i) identify the key statistical and epidemiological concepts which underlie the analysis of epidemiological data;
(ii) perform analyses of data arising from epidemiological studies, using appropriate computer software (the software used throughout will be STATA);
(iii) investigate confounding and interaction in epidemiological data;
(iv) interpret appropriately the results of these analyses, taking into account study design issues;
(iv) write a clear report presenting the results of an analysis of epidemiological data.

Our approach is to convey the concepts and methods used in the analyses and enable students to choose and use the techniques appropriate for estimation and hypothesis testing in selected situations. The course is taken by students with widely varying backgrounds, some of whom have done little statistics and others of whom are developing specialist skills in this subject. We do not delve deeply into the derivation of formulae but introduce these where we believe it will help you to see what is happening and to recognise the common threads which lie behind many of the formulations (these are brought out in Lectures 7 and 8). You do not need to know calculus.

The examples used in the sessions are all from studies in which LSHTM staff were involved. Although the topics may not be in your field, we hope that you will find that the datasets provide meaningful examples of the study designs and research questions to which our techniques apply.

The practical sessions give you an opportunity to gain experience in conducting your own analyses, and to raise individual queries. Most of you will be working in pairs in the computer sessions. We strongly recommend that you take advantage of this to discuss the results you obtain and the information you see on the screen. The interpretation of the output is as important as, if not more important than, being able to use STATA commands. Also, we need to remember that the statistical techniques are only a part of the set of tools to be used in answering research questions. The statistical results have to be set in the context of other information about the population studied, biological mechanisms, strengths and weaknesses of the study design etc.

For some of you the course may be hard but we hope that you will feel the effort is worthwhile. Many of you will then be ready to take the course in Advanced Statistical Methods for Epidemiology.

Welcome to the course.

Simon Cousens
Polly Hardy

STAFF

Study Unit Organisers

Simon Cousens
Polly Hardy

Study Unit secretary

Kirsty Ransome

Lecturers

Simon Cousens
Bianca DeStavola
Neil Alexander
Jim Todd
Craig Higgins

Practical Tutors

Ali-Akbar Haghdoost
Anjali Shah
Bernard Rachet
Bianca DeStavola
Charalambos Sismanidis
Claire Cook
Clarence Tam
Craig Higgins
Elizabeth Breeze
Elizabeth Williamson
Jim Todd
Katherine Fielding
Liam Smeeth
Mary Payne
Minouk Schoemaker
Neal Alexander
Polly Hardy
Sara Thomas
Sian Floyd
Simon Cousens
Susannah Tomkins
Tansy Edwards
Val McCormack

SME timetable 2004

Week 1				Room for Lecture
Wednesday 14 January	14.00	1	**Introduction to SME** **Summary of measures of effect (SC)**	**Goldsmiths**
	15.30		Computer practical	
Thursday 15 January	09.30	2	**Crude and stratified rates (BS)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	3	**Survival analysis (BS)**	
	15.30		Computer practical	
Friday 16 January	09.30	4	**Case control studies – concepts (SC)**	**Goldsmiths**
	11.00		Non computer practical	
	14.00	5	**Data checking/editing and univariate analysis (SF)** - Extended computer practical (runs till 5pm)	
Week 2				
Wednesday 21 January	14.00		Optional review session	Goldsmiths
Thursday 22 January	09.30	6	**Analysis of case-control studies (NA)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	7	**Likelihood theory (SC)**	**Goldsmiths**
	15.30		*Private study*	
Friday 23 January	09.30	8	**Approximate likelihoods (SC)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	9	**Logistic regression 1 (JT)**	**Goldsmiths**
	15.30		Computer practical	
Week 3				
Wednesday 28 January	14.00		Optional review session	
Thursday 29 January	09.30	10	**Logistic regression 2 (SC)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	11	**Logistic regression 3 (SC)**	**Goldsmiths**
	15.30		Computer practical	
Friday 30 January	09.30	12	**Logistic regression 4 (JT)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	13	**Matched case-control studies (SC)**	**Goldsmiths**
	15.30		Private study	
Week 4				
Wednesday 4 February	14.00		Optional review session	Goldsmiths
Thursday 5 February	09.30	14	**Analysis of matched case-control studies (NA)**	**Goldsmiths**
	11.00		Computer practical	
	14.00	15	**Strategies of analysis (CH)**	**Goldsmiths**
	15.30		*Computer practical*	
Friday 6 February	09.30	16	**Presentation of statistical results: group work**	
			Non computer practical	
	14.00	17	**Introduction to assessment**	**Goldsmiths**

SC – Simon Cousens; BS – Bianca de Stavola; NA – Neal Alexander; JT – Jim Todd;
SF – Sian Floyd; CH – Craig Higgins

Glossary of notation and terminology

As far as possible, the following standard notation and terminology will be used throughout the course.

As a general rule, Greek letters are used for **true** (population) values, while Roman letters are used for **observed** (sample) values.

Basic symbols

D,d	Number of deaths or events
H	Number of healthy people or controls
E	Expected number of deaths or events
Y	Person years at risk (pyar)
N,n	Sample size
exp(x)	$= e^x$
log()	natural logarithm = $\log_e()$
L	Likelihood
log L	Log-likelihood
E()	Expected value
Var()	Variance
SE()	Standard Error
μ	True mean ('Mu')
σ^2	True variance ('Sigma-squared')
Σ	Sum
Π	Product
P	P-value
α	Intercept in regression model ('Alpha')
β	Regression coefficient ('Beta')
y	Response variable (= 0,1)

x	Explanatory variable (exposure or confounder)
t	Time
h	Small increment of time
w	Weights used in weighted averages
z	Standard normal deviate
χ^2	Chi squared statistic
U, V	Components of a score test
Q, R	Components of a Mantel-Haenszel test statistic

Measures of disease occurrence

[Alternative terminology given in square brackets]

λ	True incidence rate ('Lambda') [incidence density]
$\hat{\lambda}$	Estimate of λ
r	Observed rate = d/Y
π	True risk or probability ('Pi') [cumulative incidence, prevalence]
Ω	True odds ('Omega')
p	Observed risk or probability
q	= 1 - p

Measures of effect

θ Rate ratio = λ_1 / λ_0 ('Theta')
[relative rate or risk]

Rate difference = $\lambda_1 - \lambda_0$

Risk ratio = π_1 / π_0 ['relative risk']

Risk difference = $\pi_1 - \pi_0$

φ Odds ratio ('Psi') = Ω_1 / Ω_0

Subscripts

1	Subscript for exposed group or case
0	Subscript for unexposed (or baseline) group, or non-case / controls
$x_1,..,x_i,..,x_I$	Explanatory variables
j = 1,...,J	Strata
k = 1,...,K	Exposure levels

Standard layout for 2 x 2 table:

	Exposed	Unexposed	
Disease +	D_1	D_0	D
Disease -	H_1	H_0	H
	N_1	N_0	N

RECOMMENDED TEXTS

There is no compulsory textbook for the Study Unit, and the course notes should provide coverage of the material. The following books are recommended for further reading:

Clayton D, Hills M. Statistical Methods in Epidemiology. Oxford University Press 1993

Kahn HA, Sempos CT. Statistical Methods in Epidemiology.Oxford University Press 1989

Kirkwood B, Sterne J. Essential Medical Statistics. Blackwell, 2nd Edition. 2003

Rothman KJ. Modern Epidemiology. Little, Brown. 1986

Rothman KJ, Greenland S. Modern Epidemiology. Lippincott Raven. 2nd Edition. 1998

Schlesselman JJ. Case-control studies. Oxford University Press. 1982

The following are advanced texts which cover the statistical methodology in greater depth than will be possible in this Study Unit.

Breslow NE, Day NE. Statistical Methods in Cancer Research. IARC Scientific Publications
Volume 1: The analysis of case-control studies. 1980
Volume 2: The analysis of cohort studies. 1987

Although the methods are illustrated with examples from cancer epidemiology, most are also directly relevant to the analysis of any disease problem.

DETAILS OF THE ASSESSMENT EXERCISE

Formal assessment of this Study Unit will consist of a report on the analysis of a data set to answer a specific research question. Each student will be asked to write a brief report summarising their approach to the analysis, the key results obtained, an interpretation of the results and their conclusions. The format will be a short written report (maximum 2 pages of A4 single-spaced) together with no more than 3 tables of appropriate summary statistics. Students should be able to show that they have understood why specific statistical methods are appropriate. Interpretation of the results and clear presentation of the findings are important.

Further details of the assessment exercise will be given in session 17. The information is delayed until then so that the assessment does not distract attention from the lectures and practicals.

The dataset to be analysed is NOT one of the datasets used during the teaching.

Two copies of the completed assessment should be handed in to Kirsty Ransome **by 4pm on Friday 13 February 2004**. Written feedback will be given to all MSc students and to others taking the SME course, if requested.

DL material

You may find the distance learning material on SME helpful. The material covered is essentially the same as in the manual, but it provides an alternative presentation and is interactive.

To access it:
1. click on School Applications, and then DBL applications. Next, click on 'Epidemiology DL'.
2. Choose the course EP202 from the menu at the top of the screen – this is SME.
3. You can then select the session you want from the course menu that appears.

DATASETS

The datasets used in this course are:

Dataset	Name	Type of study
A	WHITEHAL.DTA WHCHD.DTA	Cohort
B	MWANZA.DTA	Unmatched case-control
C	TRINMLSH.DTA	Cohort
D	DIETSME	Cohort
E	ONCHALL.DTA ONCH667B.DTA	Cross-sectional
F	DIABRAZ.DTA DIABRAZ2.DTA	Matched case-control
G	OVARSIM.DTA	Unmatched case-control

Dataset A

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. A sample of these men are included in the data set (1677 individuals).

Coding of file WHITEHAL.DTA

The coding of this dataset is given on a STATA help file set up for this course. Type **help whitehal** within STATA for this information.

Variable number	Variable name	Coding
1	id	subject
2	all	1=death from any cause; 0 otherwise
3	chd	1=death from chd; 0 otherwise
4	sbp	Systolic blood pressure at entry (mmHg)
5	chol	Cholesterol at entry (mg/dl)
6	grade4	Grade of work. 4 levels: 1=admin; 2=professional; 3=clerical; 4=other
7	smok	Smoking status: 1=never, 2=ex, 3=1-14/day, 4=15-24/day, 5=25+/day
8	agein	Age at entry (years)
9	grade	Grade of work: 1=admin/professional ("high"), 2=clerical/other ("Low")
10	cholgrp	Grouped cholesterol, 4 levels: 1= <150, 2=150-199, 3=200-249; 4 ≥250
11	sbpgrp	Grouped sbp, 4 levels: 1= <120; 2=120-139; 3=140-159; 4 ≥160
12	timein	Date of entry (days since 1/1/1960)
13	timeout	Date of exit (days since 1/1/1960)
14	timebth	Date of birth (days since 1/1/1960)

Dataset B

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.

Coding of file MWANZA.DTA

In the file MWANZA.DTA each line contains the data for one woman (case or control).
The coding of this dataset is given on a STATA help file set up for this course. Type **help mwanza** within STATA for this information.

Variable number	Variable name	Coding
1	idno	Identity number
2	comp	Community: 1-12
3	case	Case/control: 0=control 1=case
4	age1	Age group: 1=15-19, 2=20-24, 3=25-29, 4=30-34, 5=35-44, 6=45-54
5	ed	Education: 1=none/adult only, 2=1-3 years, 3=4-6 years, 4=7+years
6	eth	Ethnic group: 1=Sukuma, 2=Mkara, 3=other, 9=missing
7	rel	Religion: 1=Moslem, 2=Catholic, 3=Protestant, 4=other, 9=missing
8	msta	Marital status: 1=currently married, 2=divorced/widowed, 3=never married, 9=missing
9	bld	Blood transfusion in last 5 years: 1=no, 2=yes, 9=missing
10	inj	Injections in past 1 year: 1=none, 2=1, 3=2-4, 4=5-9, 5=10+, 9=missing
11	skin	Skin incisions or tattoos: 1=no, 2=yes, 9=missing
12	fsex	Age at first sex: 1=<15, 2=15-19, 3=20+, 4=never, 9=missing
13	npa	Number of sex partners ever: 1=0-1, 2=2-4, 3=5-9, 4=10-19, 9=missing
14	pa1	Sex partners in last year: 1=none, 2=1, 3=2, 4=3-4, 9=missing
15	usedc	Ever used a condom: 1=no, 2=yes, 9=missing
16	ud	Genital ulcer or discharge in past year: 1=no, 2=yes, 9=missing
17	ark	Perceived risk of HIV/AIDS: 1=none/slight, 2=quite likely, 3=very likely/already infected, 4=don't know
18	srk	Perceived risk of STDs: 1=none/slight, 2=quite likely, 3=very likely/already infected, 4=don't know

Dataset C

Cohort study of risk factors for mortality among males in 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 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.

Dataset TRINMLSH.DTA 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.

Coding of file TRINMLSH.DTA

The coding of this dataset is given on a STATA help file set up for this course. Type **help trinmlsh** within STATA for this information.

Variable number	Variable name	Coding
1	ethgp	Ethnic group 1=African, 2=Indian, 3=European, 4=Mixed, 5=Chinese/Semitic
2	ageent	Age in years at first survey
3	death	Died from any cause 0=no, 1=yes
4	cvdeath	Died from CV disease 0=no, 1=yes
5	alc	Drinks per week 0=none, 1=1-4, 2=5-14, 3=15+
6	smokenum	No. of cigarettes per day 0=non-smoker, 1=ex-smoker, 2=1-9, 3=10-19, 4=20-29, 5=30+
7	hdlc	HDL cholesterol
8	diabp	Diastolic BP (mmHg)
9	sysbp	Systolic BP (mmHg)
10	chdstart	Heart disease at time of entry 0=no, 1=yes
11	days	Days of follow-up
12	years	Years of follow-up
13	bmi	Body mass index (=Wt/Ht^2)
14	id	Identification number
15	timein	Date of entry (days since 1/1/1960)
16	timeout	Date of exit (days since 1/1/1960)
17	timebth	Date of birth (days since 1/1/1960)

Dataset D

Cohort study of diet risk factors and heart disease

These data arose from a pilot study of the use of a weighed diet over 7 days in epidemiological studies. The data relate subsequent incidence of coronary heart disease (CHD) to dietary energy intake.

Morris JN, Marr JW, Clayton DG
Diet and heart: a postscript
BMJ, 1977, 2, 1307-14

The data refer to the follow-up of 337 subjects, who were 30-67 years old at the beginning of the study. For each subject, an average daily total energy intake was obtained during a one week period at the start of the study. The outcome of interest was coronary heart disease (CHD) and the question of interest was the effect of total energy intake on this outcome. In particular, it was thought that a low total energy intake indicated low physical activity, so low energy intake was expected to be associated with an increased risk of CHD.

The subjects were selected from three working groups, namely bus drivers, bus conductors and bank workers. Other variables measured were height, weight and the month during which the weighed diet was obtained.

The coding of this dataset is given on a STATA help file set up for this course. Type **help dietsme2** within STATA for this information.

Coding of file DIETSME2.DTA

Variable number	Variable name	Coding
1	id	identity number
2	doe	date of entry
3	dox	date of exit
4	dob	date of birth
5	fail	code for outcome: 1 3 13 = CHD
6	job	0=driver, 1=conductor, 2=bank
7	month	month when weighed dietary survey took place
8	energy	total energy (kcals/day)
9	height	height (cm)
10	weight	weight (kg)
11	fat	total fat content of diet in g/day
12	chd	CHD indicator (1=yes, 0=no)

Dataset E

Prevalence study of onchocerciasis in Sierra Leone

Onchocerciasis (commonly known as River Blindness) is a chronic filarial infection 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 (mf) discharged by the female worm migrate through the skin, often causing intense rash and itching, dyspigmentation and atrophy. The mf 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. Mf, 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 north-east 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 mf 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 mf in the cornea or anterior chamber were recorded.

Coding of files ONCHALL.DTA and ONCH667B.DTA

There are two files. ONCHALL.DTA contains data for all 1,302 subjects and has been used to illustrate the methods of logistic regression in the notes.

ONCH667B.DTA is a reduced dataset containing data for 667 subjects who are aged >30. This restricted dataset is suitable for investigating factors affecting the prevalence of eye lesions. (Eye lesions were uncommon in persons less than 30 years old).

The coding of these datasets is given on a STATA help file set up for this course. Type **help oncho** within STATA for this information.

Variable number	Variable name	Coding
1	area	Area of residence: 0=savanna, 1=forest
2	sex	0=male, 1=female
3	agegrp	Age group (coding depends on dataset) ONCHALL.DTA: 0=5-9, 1=10-19, 2=20-39, 3=40+ ONCH667B.DTA: 0=30-39, 1=40-49, 2=50-59, 3=60+
4	mf	Micofilarial infection 0=no, 1=yes
5	mfload	Number of microfilariae in skin snip from iliac crest 0=none, 1=1-9, 2=10-49, 3=50+
6	lesions	Presence of any eye lesion 0=no, 1=yes

Dataset F

Case-control study of risk factors for infant deaths from diarrhoea

An attempt was made to ascertain all infant deaths from diarrhoea occurring over a one year period in two cities in southern Brazil, by means of weekly visits to all hospitals, coroners' services and death registries in the cities.

Whenever the underlying cause of death was considered to be diarrhoea, a physician visited the parents or guardians to collect further information about the terminal illness, and data on possible risk factors. The same data were collected for two 'control' infants. Those chosen were the nearest neighbour aged less than 1 year, and the next child in the neighbourhood aged less than 6 months. This procedure was designed to provide a control group with a similar age and socio-economic distribution to that of the cases.

Cases and controls with important perinatal risk factors were excluded from the study as follows: those with a birthweight under 1500g; twins; those with major malformations or cerebral palsy; and those whose initial stay in hospital exceeded 15 days. Also excluded were cases and controls aged under seven days, as there were very few diarrhoea deaths in this age group. During the one-year study period, data were collected on 170 cases together with their 340 controls.

In examining the risk associated with different infant feeding practices, care was taken to collect a history of the feeding mode both at the time of death (variable **milkfin**) and prior to the onset of the terminal illness (variables **milk** to **feedmode**), to allow for the possibility that the illness may have resulted in a change in feeding practice. For controls, the feeding information was collected for the same dates as their matched cases.

Description of datasets

There are two files. DIABRAZ2.DTA contains data for all 170 cases and their 340 matched controls (two controls per case).

DIABRAZ.DTA is a reduced dataset containing data for 86 cases with one matched control per case. The control is one of the two neighbourhood controls for that case, and is additionally matched to the case on age, so that their ages fall in the same broad age-group (0-2, 3-5, and 6-11 months) and differ by no more than 3 months. The remaining 84 cases were excluded because neither of the available controls satisfied the age matching criteria.

In both datasets, the mean ages of cases and controls are very similar (DIABRAZ2.DTA: cases 4.48 months, controls 4.52 months; DIABRAZ.DTA: cases 4.23 months, controls 4.34 months).

The coding of this data set is given on a STATA help file set up for this course. Type **help diabraz** within STATA for this information.

Coding of files DIABRAZ.DTA and DIABRAZ2.DTA

Variable number	Variable name	Coding
1	set	No of matched set (1-170)
2	case	1=case, 0=control
3	age	Age in months
4	agegp	Age group (months): 1=0-1, 2=2-3, 3=4-5, 4=6-8, 5=9-11
5	agegp2	1=0-2, 2=3-5, 3=6-11
6	agegp3	1=0-1, 2=2-11
7	sex	1=male, 2=female
8	bint	Birth interval (months): 1=first born, 2=<23 months, 3=24-35 months, 4=36+months
9	bwt	Birth weight (kg): 1=<2.50, 2=2.50-2.99, 3=3.00-3.49, 4=3.50+
10	bwtgp	1=≥3.00, 2=<3.00
11	meduc	Mother's education (years): 1=none, 2=1-3, 3=4-5, 4=6+
12	social	Social class: 1=underproletariat, 2=proletariat, 3=bourgeoisie
13	water	Piped water supply: 1=in house, 2=in plot, 3=none
14	wat2	1=in house/plot 2=none
15	income	Per capita monthly income (% of national minimum wage) 1=0-19, 2=20-39, 3=40-99, 4=100+
16	house	Type of house: 1=regular building, 2=shack
17	fridge	Availability of refrigerator: 1=yes, 2=no
18	milkfin	Type of milk drunk at time of death: 1=breast only, 2=breast+formula, 3=breast+cows, 4=breast+formula+cows, 5=formula only
19	milk	Type of milk drunk at onset of illness: 1=breast only, 2=breast +formula, 3=breast+cows, 4=formula only, 5=cows only
20	milkgp	1=breast only, 2=breast+other, 3=other only
21	bf	1=breastfed, 2=not breastfed
22	supp	Non-milk food supplements: 1=yes, 2=no
23	feedmode	Feeding mode: 1=exclusive BF, 2=BF+other milk, 3=other milk only, 4=BF+suppl, 5=BF+other+suppl, 6=other+suppl
24	pair	No. of matched pair (1-86) (DIABRAZ only)

Dataset G

Case -control study of risk factors for ovarian cancer

Over a 4 year period, five interviewers identified and questioned women with a diagnosis of ovarian cancer and women selected as controls at 13 hospitals in England. A standard questionnaire was used to obtain information on reproductive and menstrual history.

The study was confined to women aged less than 65 years whose diagnosis of ovarian cancer had been made within two years of interview. The analysis has been restricted to cases with a diagnosis of epithelial ovarian cancer.

The dataset **ovarsim.dta** is restricted to 203 cases and 329 controls who had been sexually active.

Coding of ovarsim.dta

The coding of the dataset is given in a Stata help file. Type `help ovarsim`.

Variable number	Variable name	Coding
1	studynumbe	Identity number
2	nulligra	Whether ever pregnant (0=ever pregnant, 1=never pregnant)
3	caco	Case-control status (0=control,1=case)
4	agegp	Age group (1=30-39, 2=40-49, 3=50-59, 4=60-64)
5	soc	Social Class (1=I&II, 2=III, 3=IV&V)
6	ocuse	Ever used oral contraceptives (0=no, 1=yes)
7	smokgp3	Smoking (0=never, 1=ex, 2=current)

STATA LICENCES

Timberlake Consultants, the distributors of STATA, operate a "graduate" scheme which allows us to buy copies of STATA at a very reasonable rate. We resell the licences at a slightly increased price to cover the cost of overheads.

Information
Purchase

1. The offer is open to currently registered PhD and MSc students and to occasional students registered for SME.

2. The offer is limited to one licence per person.

3. The cost of a STATA 8 licence is currently £100 (under the GradPlan Scheme).

4. Please use the paying-in form in this manual. Copies can also be obtained from Blackboard or from Keith Flanders in Supplies, who is located in LG5 (in the basement, Malet Street side of the main Keppel Street building). Please visit Keith to obtain your license between 2-4pm, any weekday. You will need to pay by cash or cheque – if by cheque you will need to bring your cheque card.

5. Manuals are not included in this price. These can be bought from Keith Flanders.

6. In order to run STATA on your own PC you need at least an 80836 processor (better if more advanced) and either a Pentium or Celeron chip. If your chip is 80836 or 80846 you need a maths co-processor.

7. If stocks have run out it may be necessary for the School to order more licences.

8. You will receive the CD, a Licence and authorisation key sheet, installation details and information about joining the STATA mailing list. Keep your licence sheet in a safe place as you need your licence number to install the software and if you have any problems with STATA.

9. The Stata Corporation's website www.stata.com is a very useful source of information. For example, it has a section on Frequently Asked Questions (FAQs). Although manuals can be ordered direct from the Stata Corporation this is likely to be more expensive than buying them via Stores.

Installation

The installation should be straightforward. If problems are experienced please go to Computer Advisory Service (room B116) first and if they cannot help e-mail stata@stata.com with details of the problem. You will need to tell them your licence number.

London School of Hygiene & Tropical Medicine
(University of London)

Keppel Street, London WC1E 7HT

Stata 8.0 licences
Order Form

Please take this form, with your cash or cheque, to Keith Flanders, Purchasing Officer, who is located in Room LG5, on the Malet Street side of the Keppel Street building.

Name: .....

Course or Unit: .....

Date: .....

Email .....

Amount paid: £100.00 cash / cheque * (Inclusive of Postage)
*** delete as appropriate**

PLEASE NOTE

Refunds can only be made if products are returned in a condition that makes them suitable for resale. Please ensure that you retain your receipt as you will be required to present this should you need to return the product.

LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

RECEIPT

Received, the sum of £ 100.00 for Stata 8.0 licence

From: ..... Department.....
[Full name in capitals]

Signed: ..... Date: .....
Keith Flanders

USING THE NOVELL NETWORK TO RUN STATA

For all computer practicals you should use your usual login ID

Copying the data files

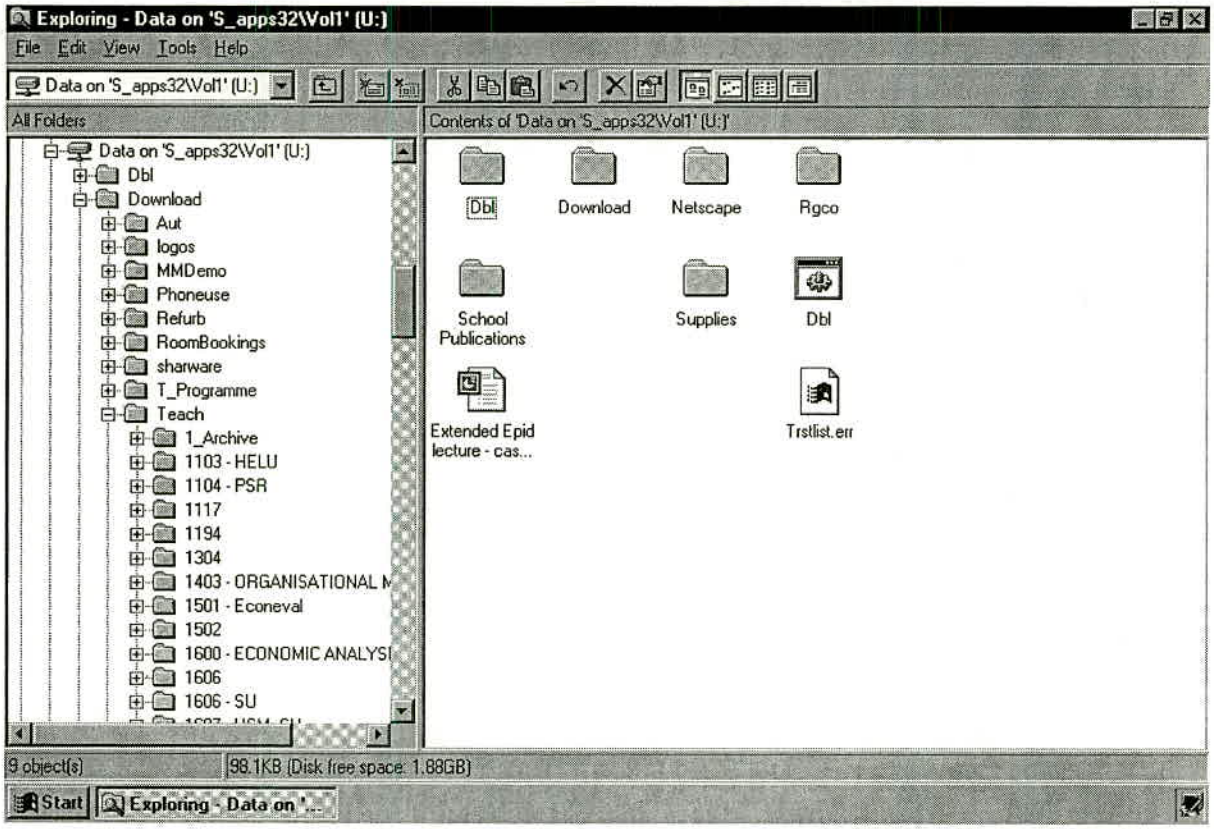
BEFORE using STATA, copy the course datasets into your home directory or personal network space (the h:\ drive). The data files are in STATA format (with the extension .dta) and are currently in the folder:-

u:\download\teach\sme

In that directory you will also find the "help files" (with the extension .hlp) These hold descriptions of the separate data files.

You should copy all the files currently in **u:\download\teach\sme** into **h:\sme**. There are many ways of doing this. If you are not familiar with copying files then carry out the following steps.

a) You can access the h:\ drive and u:\drive using Explorer which is in the 'Novell-delivered Applications' window in the 'local applications' folder. Double-click on the Explorer icon and you should get a window like the following. It is split into two panels: you may need to make the left hand one wider by putting the cursor on the dividing line and dragging it to the right.



b) In the left hand panel (**All Folders**), scroll down and select :-

Data on 'S_apps32\Vol1'(u:\download\teach)

c) In the right hand panel you will see a folder called **sme.** Select this folder (click the right mouse button just once on the folder name - **sme**) and select Copy.

d) In the left panel scroll up to the h:\ drive, and double-click on it. Click the right mouse button on the right hand panel and select Paste. The files should now be in your "home space" (h:\ drive) in the folder called **sme**.

If you accidentally delete a data file during the course you can copy it again in this way from the **u\:** drive.

Using STATA

1) In Windows NT click on the STATA8 icon in the LSHTM Statistical Applications Folder

2) In the Command box type:- **cd h:\sme**

This ensures that you are working in the subdirectory (or folder) in which your data files are kept. All your logs and new data files will be together in this subdirectory (or folder).

3) Log files

a) To keep a record of the commands you use and output you generate during the session, type the following syntax in the Command box:-

log using <logname>, text replace

Insert whatever name you choose to call the log for that session, eg sme1 sme2 etc in place of **<logname>**. You only need the replace option if you already have a file called **<logname>** that you want to replace. If you are creating a file for the first time, you do not need the replace option.

The log file will have the extension **.log**

To close the log file, type:-

log close

You can look at your log at any point during the session by clicking on the **File** menu, and selecting **Log**, and then **View**.

b) To keep a record of the commands that you use, but not the output in the Command box type:-

cmdlog using <logname>, replace

To close this file, type

cmdlog close

This is useful for saving a list of commands that you can later edit into a **do** file (see below and Practical 1). Do files are helpful because once you have worked out the right sequence of commands and you want to do a similar analysis again, you don't need to keep working it all out afresh.

By default, the log file you create using the command **cmdlog** will have the extension **.txt**. If you wish to change this, you must type the extension. For example, to create a log file called sme1.do (to indicate that it is to be used as a do file), type:-

cmdlog using sme1.do, replace

You can create the files described in a) and b) simultaneously.

c) If you wish to edit your log file after the session, you can open it in WORD. Please EDIT OUT all duplications and mistakes before printing your log. Otherwise you will be wasting paper with output of no use to you and keeping your co-students waiting for the printer.

4) DO files

You can write and save your STATA commands in a text editor within STATA called the do-file editor. Click on the icon in STATA that looks like an envelope with a pen on it (8th button from the left) to open the do-file editor.

A simple do file may look like this:-

```
*
*       session1.do       (include comments by starting the line with an asterisk)
*
*       Do file to perform the exercises for SME practical session 1
*

capture log close                          (this closes any log file already open)
cd h:\sme                                  (changes default directory to h:\sme)
log using session1.log, replace            (opens log file to go with this do-file)
use whitehal,clear                         (opens the dataset you wish to use)
```

You can run individual commands or sequences of commands from the do-file editor by selecting/highlighting the commands you want to run, clicking on **Tools** and then clicking on **Do selection**.

HOW TO USE THIS GLOSSARY

This glossary provides a listing of STATA commands used or implied by the practicals in SME and ASME study units. 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.

1. Utilities
2. Data manipulation and management
3. Descriptive statistics
4. General statistics
5. Cohort/survival analysis
6. Case-control/cross-sectional analysis
7. Regression models
8. Likelihood exercises

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.

GENERAL INFORMATION ABOUT STATA COMMANDS

Syntax

command <varname(s)> if .... in..... , options

or

by <varname>: command <varname(s)> if ... in ..... , options

Help

Use **help** or **lookup** command to find out more about any commands

In-house ado files

Some of the commands introduced in SME and ASME have been developed at LSHTM. To use them with your personal copy of STATA, you will need to copy them into the folder of C:\ADO of your own pc.

1. Utilities

clear	Clear data from memory
display	Display values from functions
do	Execute commands from a file
exit	Exit STATA
help	Obtain on-line help
log	Record all commands and output entered during session into a file
lookup	Obtain on-line help; useful if the term you use is not a command term
save	Save data into a STATA file
use	Use a STATA data set

2. 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, e.g. number of decimal places to display or how date will appear
generate	Create new variable
infile	Read non-STATA data into memory, e.g. ASCII text file
input	Enter data from keyboard
keep	Keep a subset of variables, e.g. for new dataset
label	Manipulation of labels for variables, categories or datasets
list	List values of variables
merge	Merge two sorted data sets
mvdecode	Change a numeric code for a missing value to the STATA-defined for 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

3. Descriptive statistics

list	List values of variables
graph twoway scatter	Display observations graphically
histogram	Histogram of categorical variable
summarize	Display summary statistics
table	Multiple two-way tables of frequencies and summary statistics
tabulate	One- and two- way tables of frequencies

4. General statistics

anova	Analysis of variance
correlate	Correlation between variables
oneway	One-way analysis of variance
ranksum	Wilcoxon ranksum test
table	Multiple-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

5. 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

6. 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)

7. Regression models

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)
estimates store	Stores logistic regression parameter estimates
lrtest	Likelihood ratio test
poisson	Poisson regression (output on log scale)
regress	Linear regression

xi: command i.varnames	Allows you to fit a regression model where categorical explanatory variables are included as a set of indicators

8. Likelihood simulations

blik	Binomial likelihood curve
bloglik	Binomial log likelihood curve
plik	Poisson likelihood curve
ploglik	Poisson log likelihood curve

STATISTICAL METHODS IN EPIDEMIOLOGY

INTRODUCTION TO STATA 8.1

This practical is intended primarily for students who are not very familiar with STATA. We suggest that you do it in your own time. You should do the main practical on the first afternoon; try to share with someone familiar with STATA while doing the main practical.

Objectives:

Having worked through this practical students will be able to:

i) know the different components of STATA (dataset, log, command, output, and the purpose of different windows obtained in stata);
ii) get into STATA, open a dataset and a log and begin analysis;
iii) obtain data descriptions;
iv) produce crosstabulations;
v) create a new variable derived from information already contained in existing variables;
vi) produce some straightforward graphs;

PART I

1.1 The STATA package

STATA is a general purpose statistical package which works on data files stored entirely in memory (RAM). For this reason it is very fast. It also has excellent data handling and graphics facilities.

STATA is a command language. A typical command is

```
tabulate smok grade if agein>50, col
```

which tabulates the two categorical variables **smok** and **grade** for all records for which age at entry (**agein**) is greater than 50. It also uses the option **col** which sets out the table with column percentages.

The general form of STATA commands is

```
command varname(s) if ... in ... using ... , options
```

where **if** and **in** lead to select a subset of records on which the command is executed, **using** calls other data files and **options** are features specific to each command. The general style of STATA is to give minimal output as standard, and to leave extras to be specified as options.

STATA distinguishes between upper and lower case. You must always use lower case when typing commands, and we recommend that, in general, you also use lower case for variable names. STATA accepts abbreviations for commands and variable names providing they are not ambiguous. While in STATA, you can press the PAGE UP or PAGE DOWN to recall and edit previous commands. Alternatively one can copy and paste from the Review Window (see Section 1.3).

1.2 The data

To introduce you to STATA we use the Whitehall data **whitehal.dta**, a subset of a large study of risk factors for ischaemic heart disease (IHD) carried out between 1967-69 on male civil servants from various departments around Whitehall (London). The variables in this study are shown in Table 1. The first 10 records of a selection of variables are shown in Table 2.

Variable	Unit or coding	Type	STATA name
Subject number	---	---	**id**
Death indicator	1=death from any cause; 0=alive	Binary	**all**
CHD death indicator	1=death from CHD; 0=otherwise	Binary	**chd**
Systolic blood pressure at entry	mmHg	Quantitative	**sbp**
Cholesterol at entry	mmol	Quantitative	**chol**
Grade of work (in 4 categories)	1=admin; 2=professional; 3=clerical; 4 =other	Categorical	**grade4**
Smoking status	1=never, 2=ex, 3=1-14 cigs/day 4=15-24 cigs/day 5=25+ cigs/day	Categorical	**smok**
Age at entry	years	Quantitative	**agein**
Grade of work (in 2 categories)	1=admin/professional; 2= clerical /other	Binary	**grade**
Grouped cholesterol (in 4 categories)	1= ≤149; 2=150-199; 3=200-249; 4 =250+	Categorical	**cholgrp**
Grouped sbp (in 4 categories)	1= ≤119; 2=120-139; 3=140-159; 4 =160+	Categorical	**sbpgrp**
Date of entry	days since 1/1/1960	Date	**timein**
Date of exit	days since 1/1/1960	Date	**timeout**
Date of birth	days since 1/1/1960	Date	**timebth**

Table 1: Description of **whitehal.dta**

id	all	sbp	chol	Grade	smok
5001	0	120	273	2	4
5019	0	118	234	1	3
5038	0	147	295	1	3
5039	0	92	210	1	1
5042	1	128	287	1	2
5052	1	109	209	1	2
5064	1	145	262	1	1
5078	1	144	268	1	2
5089	0	154	187	1	1
5090	0	114	222	1	1

Table 2: List of a selection of variables for the first 10 records of **whitehal.dta**

Remember that categorical variables record into which category a subject falls. The different categories are often coded using numbers, and they may sometimes have a logical order, as in: poor, medium, good. When there are only two categories the categorical variable is called binary. Quantitative variables record a measurement or count of some kind.

1.3 Starting STATA

To connect to the School Network, see the instructions in the Course Handbook at the beginning of the manual. To open STATA8, find the folder called "Statistical Applications", click on it and then on the STATA8 icon.
You will see four STATA windows below two lines of menu buttons:

1. the **Command** window where you type the commands;
2. the **Results** window where you read the output;
3. the **Variables** window where the variables read by STATA are listed;
4. the **Review** window where everything you type in the **Command** window is listed.

1.4 Keeping a log

It is good practice to keep a record of every STATA session by typing in the **Command** window,

```
log using smela, text
```

We suggest the name **smela** *but you could use any convenient name for your log file. Note that the actual file produced by the* **log** *command had the extension* **.log** *(i.e. the full name is* **smela.log**).

If you wish to temporarily stop recording your output, type log off. Type log on to start recording your output again.

When you want to close your log file, type

```
log close
```

1.5 Reading the data

The file **whitehal.dta** contains the variable names and values for the 1677 records of the subset of the Whitehall study and should be in your local directory **h:\sme** if you have followed the instructions on page xxiii of the introductory part of the manual to copy data sets from a central directory to yours. To read the data, type

```
use whitehal
```

The full name of this file is **whitehal.dta** *but there is no need to include the extension* **.dta** *since STATA assumes all files read in with* **use** *have the extension* **.dta**.

If you now type **describe** in the **Command** window you will obtain the description of the data held in memory, that is **whitehal.dta,** including names and types of all the variables. (The types refer to how the values are stored, whether as bytes, integers, long integers, or floating point numbers. The distinction between these can be important when you are analysing large data sets, but don't worry about it now.)

If you type

```
help whitehal
```

more information on this data set will appear on the screen.

1.6 A first look at the data

A good way to start is to ask for a summary of the data by typing

```
summarise
```

This will produce the mean, standard deviation, and range, for each variable in turn.

> *STATA cannot distinguish between categorical variables, which are coded using numbers, and quantitative variables, so it produces mean and standard deviations for both. When categorical variables are coded alphabetically (for example if* **grade4** *had been entered as a string variable with values "admin", "professional", "clerical" and "other" instead of their numerical codes) the mean and standard deviations are not produced.*

In most datasets there will be some missing values. These are usually coded using the symbol . (i.e. "full stop") in place of the value which is missing. STATA can recognize other codes for missing values, but this is the one which is recommended. The **summarize** command is useful for seeing whether there are missing values.

The **list** command is used to list the values in the data file. Try out the following and see their consequences:

```
list in 1/5
list agein in 1/10
list grade sbp
```

> *STATA stops after each screenful of output. Press* ENTER *to continue line by line, and press* SPACE *to get another screenful. The command* **list** *on its own would list all of the data. You can cancel this command (and any other STATA command) by pressing* CTRL-BREAK *(break is the same key as* PAUSE*). Alternatively, you can cancel any command by clicking on the red symbol with a white cross on it, on the far right of the toolbar in STATA.*

1.7 Cross-tabulating

When starting to look at any new data the first step is to check that the values of the variables make sense and correspond to the codes defined in the coding schedule. For categorical variables this can be done by looking at one-way frequency tables and checking that only the specified codes occur. For quantitative variables we need to look at ranges and histograms.

This first look at the data will also indicate whether all values are present or whether there are some missing values on some variables. Let us begin by looking at the categorical variables. The distribution of the categorical variables **smok** and **grade** can be viewed by typing

```
tab smok
tab grade
```

Their cross-tabulation is obtained by typing

```
tab smok grade
```

Cross tabulations are useful when checking for consistency. The basic output from a cross tabulation reports frequencies only; to include row and/or column percentages add the options `row`, `col`, `cell`, or any combination, as in

```
tab smok grade, col
```

The `tab` command can be used with any type of variable provided the number of different values is not too large. STATA allows a maximum of 500 different values, but the result will not be useful for variables with as many different values as this. To demonstrate this, try

```
tab sbp
```

Optional
The smoking variable is called `smok` but it would be preferable if it were labelled more clearly. We can do this by typing,

```
label var smok "Smoking status at entry"
```

Also, to recall its coding, we could use the following set of commands,

```
label define labsmok 1 "never" 2 "ex" 3 "1-14 cig"  4 "15-24 cig"  5 "25+
cig"
label value smok labsmok
```

which define the numerical codes of `smok` using the indicator `labsmok`. If you then also label the variable `grade` in the same fashion,

```
label var grade "Grade of work at entry"
label define labgrade 1 "high" 2 "low"
label value grade labgrade
```

the output will be much improved when you will type again `tab smok grade`.

1.8 More details

The command **summarize** has already been used to give means and standard deviations for all variables together. To describe a particular variable such as **sbp** in more detail, type

```
summarize sbp, detail
```

This command will give the mean, the standard deviation, the median, a selection of percentiles and the minimum and maximum values of **sbp**. *It will also give the skewness and kurtosis of the distribution of* **sbp**, *quantities which are rarely of much interest.*

To obtain the means, standard deviations and medians of **sbp** separately by **grade**, type

```
table grade, c(freq mean sbp sd sbp median sbp) row
```

The **c( )** *option in the* **table** *command stands for ``contents'' and allows several entries besides those used here: for example* **p25 sbp** *would compute the 1st quartile of* **sbp** *and* **p75 sbp** *the third. The* **row** *option generates the total row at the bottom of the table.*

1.9 Restricting commands

STATA commands can be restricted to records 1, 2, ..., 10 (for example), by adding " **in 1/10** " to the command. The letters **f** and **l** can also be used as abbreviations for "first" and "last", so **20/l** refers to the records from 20 onwards. Commands can also be restricted to operate only on records which satisfy given conditions. The conditions are added to the command using **if** followed by a logical expression which takes the values true or false. For example, to restrict the command list to records with **sbp** greater or equal to 120 mmHg, type

```
list if sbp>=120
```

If the logical expression **sbp>=120** *is true the record is listed, but not otherwise. Other useful logical expressions are* **==** *for "equal", and* **!=** *for "not equal".*

A useful command when exploring data is **count** which counts the number of records which satisfy some logical expression. For example

```
count if sbp>=120
count if sbp>=120 & smok==1
```

Note the use of **&** *to link two conditions both of which must be satisfied. Also note that a common error is to use* **=** *in a logical expression instead of* **==**.

1.10 Generating and recoding variables

New variables are generated using the command **generate**, and variables can be recoded using **recode**. For example, to create a new variable **higrade** which is the same as **grade** but coded 1 for "High grade" and 0 for "Low grade" (instead of 1 for "High grade" and 2 for "Low grade"), try

```
gen higrade=grade
recode higrade 2=0
tab higrade grade
```

To group the values of a metric variable, such as **sbp**, we use **egen** and **cut**. For example,

```
egen sbpcat=cut(sbp), group(4) label
table sbpcat, c(freq min sbp max sbp)
```

creates a new variable, **sbpcat**, which takes the value 0 when **sbp** is in the first quartile, 1 when **sbp** is in the second quartile, etc. The option **label** at the end of the command makes sure that the cut-points defining the quartiles are reported every time the variable is tabulated.

1.11 Sorting

The records in a dataset can be sorted according to the values of one or more variables. The **whitehal** dataset is currently sorted by **id** but for some purposes it might be better to have it sorted by **agein**. Try

```
sort agein
list id agein
```

The records are now in order of **agein** and **id** as well as all other variables are sorted in this order.

> *Note that STATA commands which use the option* **by( )** *require the data to be first sorted by the variable in the* **by( )** *option (the sorting is not done automatically because you should always be aware of how your data are held).*

1.12 Saving the data

In order to keep these new variables for use in future sessions type

```
save whnew
```

> *This will save the data currently in memory in the file* **whnew.dta**.

If you already have a copy of **whnew.dta** but you wish to replace it with a new version just type,

```
save whnew, replace
```

PART II

2.1 Graphical displays

The **graph** command has many options. Bar charts are used to display the distributions of categorical variables, while histograms and box plots are used to display the distributions of quantitative variables.

To obtain a histogram of **sbp**, type

```
histogram sbp
```

You can vary the number of rectangles in the histogram (called bins) by adding the **bin( )** option where, for example, **bin(8)** would lead to 8, potential, rectangles (there could be less than 8 if the data are too sparse; 50 is the maximum number allowed for this option). To superimpose the histogram with a normal curve which has the same mean and standard deviation as the data, add the option **normal.** Try, for example,

```
histogram sbp, bin(8) normal
```

2.2 Box plots

An alternative to the histogram is the box plot, which is obtained using **graph box**. For example, try

```
graph box sbp
```

2.3 Multiple graphs

To put several box plots in a graph, for example to examine the distribution of **sbp** by **grade**, try

```
sort grade
graph box sbp, by(grade)
```

2.4 Scatter graphs

Scatter plots can be used to evaluate the association between **sbp** and the other quantitative variables in the data set. For example,

```
graph twoway scatter sbp chol
graph twoway scatter sbp agein
```

show that **sbp** is associated with **agein.**

To combine more than one scatter plot in the same graph it is best to use the **by()** option. For example, to show the scatter plot of **sbp** against **agein** separately by **grade**, with all plots on the same graph, type

```
sort grade
graph twoway scatter sbp agein, by(grade)
```

2.5 Help

Whenever you want more information on a command , or you have forgotten its syntax, you can type

```
help <command name>
```

and STATA will produce a detailed listing. See for example, **help use** and **help graph** (if you wish to interrupt the listing use ***CTRL-BREAK***).

There is also a help file for all the data sets used in this course, e.g. try **help whitehal** .

2.6 Leaving STATA

To leave STATA when you have completed the analysis you need to type,

```
exit, clear
```

where , **clear** is necessary because the data are still in memory and STATA is careful about warning when you might lose data by mistake.

If you only wish to clear the data in memory without leaving STATA just type **clear**.

FURTHER EXERCISES

1. List the variables **id**, **chol** and **smok** for records 20-25 inclusive.

2. Obtain the frequency distribution of **chol** using **histogram** (try different values for **bin ()**).

3. Obtain the two way table of frequencies of **smok** and, **sbpgrp.** Use in turn the options: **row, col** and **cell.**

4. Use **count** to find how many members of the cohort had **sbp>200.** Then see how many among them were High grade (use **& grade==1**).

5. Create a new variable called **agecat**, by grouping the values of **agein** into (30-49, 50-54, 55-59, 60+) using
   egen agecat=cut(agein), at(30,50,55,60,90)
   Check the possible options of this command using **help cut**.

6. Use

   table agecat, c(min agein max agein)

   to check that you have defined **agecat** appropriately.

Statistical Methods in Epidemiology (2402)

Course Manual 2004

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 1

Measures of disease occurrence and exposure effects

Objectives

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

(i) state which measures of effect can be used with each of three common epidemiological study designs
(ii) describe the relationship between the risk, rate and odds and their respective ratio measurements
(iii) list the advantages of rates (and rate ratios) as a measure of effect
(iv) explain why the odds ratio rather than the risk ratio is commonly used as the effect measure in the analysis of case-control and cross-sectional studies

1. Introduction

Many epidemiological studies are performed to address questions of the form:

Does the chance of infection/disease/mortality change with the different levels of some factor (exposure) even after controlling for potentially confounding variables?

Statistical techniques enable us to respond to such questions and to quantify the uncertainty in our response due to random variation.

Most epidemiological studies can be analysed in several different ways and, conversely, different epidemiological study designs can be analysed using some of the same statistical methods. In this course we focus on the application of statistical techniques to the analysis of three common epidemiological study designs. These are follow-up (cohort) studies, case-control studies and cross-sectional studies. We start by looking briefly at examples of each type of study that we will study in more detail as the course progresses.

Example 1: The Whitehall study (dataset A).

This study followed-up civil servants for 20 years. Subjects were recruited into the study over a two-year period, and mortality was recorded over the follow-up period. One exposure of interest was the grade of work, classified as high (administrative, professional or executive) and low (clerical or other).

Example 2: Case-control study of HIV infection in women (dataset B).

Cases were women found to be HIV-positive in a cross sectional survey of 12 communities in Mwanza, Tanzania. Controls were randomly selected from women

who were HIV negative. Risk factors of interest included educational level of the women, and sexual practices including number of sexual partners.

Example 3: Cross-sectional study of onchocerciasis (dataset E)

The prevalence of onchocercal infection and eye disease among villagers living near river breeding sites for blackfly were measured in a cross-sectional study. Of interest was the relationship between the ecological zone in which the communities were living and the prevalence of infection with *Onchocerca volvulus*.

2. Measures of disease outcome

Different types of study lead to the use of different outcome measures:

- for **cohort** studies, the outcome is usually the **risk** or **rate** of disease.
- for **cross-sectional** (prevalence) studies, the outcome is usually the **risk** (prevalence) of disease or **odds** of disease.
- for **case-control** studies, we cannot estimate the absolute risk, rate or odds of disease, but we can estimate the **odds ratio**.

In this session we shall consider comparisons between two groups of individuals, an **exposed** group who possess the risk factor of interest, and an **unexposed** group who do not. We are interested in comparing the amount of infection/disease/mortality in the exposed group to that in the unexposed group.

For all three outcome measures (risk, rate, odds) and their ratios, we assume that there is some **true** or **underlying** value for the population from which the data were sampled. We wish to **estimate** this value, using the data from our study. We shall usually also wish to calculate a confidence interval which gives an idea of the likely range within which the true (population) value lies, and we may wish to test whether the true value differs from some hypothesised "null" value. Methods for calculating confidence intervals and performing hypothesis tests will be discussed later in the course.

3. Risk measures

Consider a group of N disease-free individuals who are recruited into a follow-up study at a particular point in time and are then all followed for a fixed length of time. The following table shows the outcomes:

Disease	Exposed	Unexposed	Total
Yes	D_1	D_0	D
No	H_1	H_0	H
Total	N_1	N_0	N

The risk of incident disease is defined as the probability that an individual experiences the disease during the period of follow-up. Concentrating first on exposed individuals, we can estimate the risk of disease as

$$\pi_1 = D_1 / N_1$$

We can obtain a similar measure for unexposed individuals; the risk (π_0) of disease in the unexposed group is estimated as D_0 / N_0.

The effect of an exposure on disease is usually assessed by taking the **ratio** of disease frequency in exposed individuals to that in unexposed individuals. (But be careful – "effect" carries the implication of causation. In observational epidemiology we measure associations rather than effects.) Using this approach, we obtain

Risk ratio = π_1/π_0

Another measure of exposure effect which is sometimes used is the risk **difference** — the difference between the risk of disease in exposed and unexposed individuals:

Risk difference = $\pi_1 - \pi_0$

Exercise 1: As an example, consider a group of 30,000 subjects observed for 10 years, after which time 50 were diseased. The results might be tabulated as follows:

Disease	Exposed	Unexposed	Total
Yes	30	20	50
No	9 970	19980	29 950
Total	10 000	20 000	30 000

i) Calculate the risk of disease separately for exposed and unexposed individuals and obtain the risk ratio for exposed vs unexposed individuals.

Risk in exposed = Risk in unexposed = Risk ratio =

ii) Calculate the risk difference for exposed vs unexposed individuals.

Risk difference =

4. Rate measures

The risk of disease tends to increase with the length of the time interval considered: the longer an individual is observed the greater their chance of developing the disease. It follows that a 5 year risk can only be compared with other 5 year risks; not 10 year risks, etc. Some further disadvantages of risks are

1. They assume that all individuals are followed for the same length of time. They are therefore unsuitable for studying cohorts where individuals may enter or exit at various points during the period of observation. An example is occupational studies where individuals may join (or leave) the workforce at various times during the defined period of observation. Similarly, they do not take account of individuals who are lost to follow-up (e.g. through emigration) or who cease to be at risk of a particular disease (e.g. because they have died from some other cause).

2. For individuals who develop the disease, no account is taken of **when** disease onset occurred.

3. They are not easily applied in situations in which exposures change with time, e.g. where individuals who are initially unexposed become exposed at some point during the follow-up period. An important example is the effect of age, since individuals inevitably increase in age with increasing follow-up.

Most of the disadvantages of risks are overcome by measuring disease occurrence using **rates**.

5. Estimation of disease rates

The **rate of disease** (λ) for an individual is their instantaneous risk of disease per unit of time. We estimate disease rates by observing a **group** of individuals who are assumed to have the same disease rate λ. Examples of such groups might be individuals in a particular age group who are the same sex and living in a particular geographical area. Each individual contributes a certain amount of observation time. An individual's observation time starts when the subject joins the study and stops when the subject develops the disease of interest, or is lost to follow-up, or the follow-up period ends, whichever happens first. The observation time is therefore the time during which, were the subject to experience an event, the event would be recorded in the study.

To estimate the rate (λ), we:

i) calculate the total number of events observed among all individuals, D
ii) calculate the sum of the individual observation times, Y
iii) calculate $\lambda = D/Y$.

The total observation time, Y, is called the **person-time at risk**. If it is measured in years then it is called the **person-years-at-risk**. The estimate of λ obtained in this way is called the **incidence rate**.

Exercise 2: In the Whitehall study of mortality among British civil servants, individuals were classified as never or ex-smokers, and current smokers. Complete the table (which shows the total deaths D and person-years Y in each group) by calculating the rate per 1000 person-years in each group.

	D	Y	Rate/1000 person years
Never/ex smokers	359	32145.4	
Current smokers	484	22713.3	

6. Relationship between risks and rates

For a sufficiently short interval of time, h, and low rate, λ, the risk of an event π is approximated by λ x h i.e.:

$$\text{Risk} = \text{Rate} \times \text{Time}$$

Thus, if $\lambda = 0.00164$ per day, the risk of an event occurring in one week is simply the rate per day times 7, i.e. $0.00164 \times 7 = 0.01148$. However, the risk of an event occurring during a longer period is **not** simply λh. For example, if $\lambda = 0.6$ per year, the risk of an event occurring during a five-year period is clearly **not** $0.6 \times 5 = 3.0$, since risks are probability measures and cannot exceed 1. In order to obtain risks for longer periods we must use the following formula (see Appendix 1 for derivation):

$$\text{Risk} = 1 - \exp(-\text{Rate} \times \text{Time})$$

where exp(x) stands for the mathematical function e^x. Thus if $\lambda = 0.6$ per year, the risk of an event occurring in a five-year period is

$$1 - \exp(-0.6 \times 5) = 0.95.$$

Note that in the above formula

$$\text{Risk} = 1 - \exp(-\text{Rate} \times \text{Time}) = \text{Rate} \times \text{Time}$$

if either λ, the disease rate is small (rare disease) or the follow-up time is short.

Exercise 3: The table below shows an individual's risk of developing the disease over a one year and a five year period, for different rates of disease. Complete it by calculating the risk over a 5 year period if the rate is 0.2 per year, and if rate = 0.002 per year.

	Rate=0.2 per year	Rate=0.02 per year	Rate=0.002 per year
Risk in 1 year period	0.181	0.0198	0.002
Risk in 5 year period		0.0952	

7. Definition of effect measures based on rates

Just as we compared disease occurrence in exposed and unexposed individuals by taking the ratio of their disease risks, we can also take the ratio of their disease rates.

Rate ratio = λ_1/λ_0

where λ_1 and λ_0 are the disease rates in exposed and unexposed individuals respectively. Similarly, the rate difference for exposed and unexposed individuals is obtained as

Rate difference = $\lambda_1 - \lambda_0$

Exercise 4: Calculate the rate ratio and rate difference for smokers (exposed) compared to never/ex-smokers in the Whitehall study. (See Exercise 2).

Rate ratio =

Rate difference = per 1000 person-years

8. Odds of disease, and odds ratios

In cohort studies, we have seen that outcome measures based on rates are generally preferred to those based on risks. In prevalence (cross-sectional) studies and case-control studies, subjects are seen at only one point in time, so that rates of disease cannot be calculated. The results of such studies can be tabulated as:

Disease	Exposed	Unexposed	Total
Yes	D_1	D_0	D
No	H_1	H_0	H
Total	N_1	N_0	N

We have already seen how to calculate risks, risk ratios and risk differences from such tables. However the analysis of such studies (particularly case-control studies) is usually based on **odds ratios**.

The **odds** of disease is defined as the probability that an individual experiences the disease divided by the probability that they do not:

$$\Omega = \pi/(1-\pi)$$

In exposed individuals: $\pi_1 = D_1/N_1$, and $1-\pi_1 = H_1/N_1$ (since $N_1=D_1+H_1$).

The odds in exposed individuals is therefore estimated as

$$\Omega_1 = \frac{D_1/N_1}{H_1/N_1} = D_1/H_1$$

Similarly, the odds in unexposed individuals is estimated as

$\Omega_0 = D_0/H_0$.

The **odds ratio** is defined as

$\Psi = \pi_1/(1-\pi_1)$ divided by $\pi_0/(1-\pi_0)$.

This is estimated as $\frac{D_1/H_1}{D_0/H_0} = \frac{D_1 \times H_0}{D_0 \times H_1}$

Note that if you know the risk, you can always work out the odds, and vice versa. This is because

$\Omega = \pi/(1-\pi)$, so that $\Omega \times (1-\pi) = \pi$

$\Omega - \pi\Omega = \pi$

$\Omega = \pi(1+\Omega)$

and so $\pi = \Omega/(1+\Omega)$

Exercise 5:

i) Calculate the odds of disease, if the risk is

(a) 0.01 (b) 0.1 (c) 0.5 (d) 0.8

ii) Calculate the risk of disease, if the odds are

(a) 0.05 (b) 1 (c) 2 (d) 99

9. Odds ratio as an approximation to the risk ratio

For a rare disease,

$$\pi_1/(1-\pi_1) \approx \pi_1 \quad \text{and} \quad \pi_0/(1-\pi_0) \approx \pi_0$$

so the ratio of odds is approximately equal to the ratio of risks. The more common the disease, the less close will be the agreement between the risk ratio and the odds ratio. In these circumstances, the odds ratio will be more extreme (further from 1) than the risk ratio.

Exercise 6: Imagine a fixed population of 50,000 individuals, 20% of whom are exposed to a hazardous agent. Data on disease status are obtained for exposed and unexposed individuals at ten and twenty years following exposure:

Disease	Exposed	Unexposed
After 10 years		
Yes	300	600
No	9 700	39 400
After 20 years		
Yes	2 000	4 000
No	8 000	36 000
Total	10 000	40 000

Calculate separately for the data obtained at each point in time:

(i) the risk ratio for exposed vs unexposed

After 10 years: After 20 years:

(ii) the odds ratio for exposed vs unexposed

After 10 years: After 20 years:

The analysis of case-control and prevalence (cross-sectional) studies is usually based on odds and odds ratios rather than risks and risk ratios. There are two main reasons for this:

(i) In case-control studies, we cannot calculate the risk of disease, because we do not know the total number of cases and non-cases in the population. However, we can estimate the odds of exposure, making it possible to estimate the odds ratio. This will be discussed further in the session on case-control studies.

(ii) Statistical methods based on odds ratios have better mathematical properties than those based on risk ratios. Risks are constrained to be between 0 and 1 and log(risk) to be between $-\infty$ and 0. Odds can take any value between 0 and ∞, hence the log(odds) can take any value between $-\infty$ and ∞. It is easier to

model a quantity that is unconstrained than one which is constrained. As we shall see later in the course, logistic regression is based on modelling log(odds) and provides estimates of odds ratios.

So: If the risk of disease is small then the odds ratio is approximately the same as the risk ratio. If the risk of disease is not small then the odds ratio has better mathematical properties but is a less intuitive measure.

10. Rate ratios, risk ratios and odds ratios

We saw in section 6 that, for rare diseases (small λ),

$$\text{Risk} = 1 - \exp(-\text{Rate} \times \text{Time}) \approx \text{Rate} \times \text{Time}$$

It follows that for rare diseases (or short follow-up times)

$$\text{Risk ratio} \approx \text{Rate ratio}.$$

But also from section 9 it follows that for rare diseases (or short follow-up times)

$$\text{Risk ratio} \approx \text{Odds ratio} \approx \text{Rate ratio}.$$

If the disease is not rare, for example when the rates in exposed and unexposed groups are constant but the duration of follow-up increases, the odds ratio will tend to be more extreme (further from 1) than the rate ratio while the risk ratio will be less extreme than the rate ratio.

Exercise 7: Consider the following data obtained from a follow-up study of 10,000 individuals observed for a period of 2 years. Calculate the odds ratio, risk ratio and rate ratio and compare them.

Disease	Exposed	Unexposed
Yes	15	30
No	1 985	7 970
Total	2 000	8 000
Pyears	3 985	15 970

Odds ratio = Risk ratio = Rate ratio =

11. Measures of exposure effect: ratios versus differences

We have mentioned two ways of assessing the effect of exposure on disease — ratio measures and difference measures. In general, **ratio measures** tell us about the **strength of association** between exposure and disease and are of central importance

to studying the aetiology of disease. **Difference measures** (risk or rate differences) are generally held to be useful for **assessing the public health implications** of an exposure.

Example: Mortality of British male physicians 1951-1961

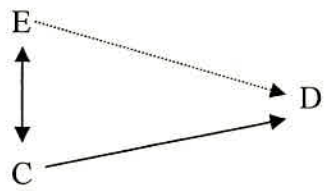
	Death rate per 1000 person years		Rate	Rate
	Non smokers	Smokers	ratio	difference
Lung cancer	0.07	2.27	32.4	2.20
Cardiovascular	7.32	9.93	1.4	2.61

The importance of smoking as an aetiological factor for lung cancer is clearly demonstrated by the large rate ratio (RR=32.4). In contrast, smoking appears less important in causing cardiovascular disease (RR=1.4). From a public health viewpoint, however, smoking is just as important for cardiovascular disease as lung cancer because the absolute increase in death rate is similar for both diseases.

The most commonly used statistical models estimate ratios (rate, odds), making the default assumption that the effects of different risk factors combine multiplicatively.

12. Confounding and effect modification

In epidemiological studies we generally observe associations but are often interested in making inferences about causation. Bradford-Hill has provided criteria to assist in judging whether an association is causal. We would like to rule out as far as possible other explanations for observed associations, one of which is confounding. Suppose we are interested in the association between exposure E and disease D. Confounding occurs when there is some other factor C, which is itself associated with D and is also associated with E.



Failure to take C into account when examining the association between E and D will produce misleading results; our estimated effect measure (risk/rate/odds ratio) will incorporate not only the association between E and D but also, to some extent, the association between C and D.

Effect modification (interaction) is said to occur when the effect or exposure (eg risk/rate/odds ratio) varies according to the level of some other factor.

We use the same approach, stratification, to address both confounding and effect modification, but it is important to understand that they are different phenomena.

Appendix 1: Derivation of the relationship between risk and rate

Consider one individual at a particular point in time. The **risk** that the individual experiences an event in the next short interval of time depends on the length of that time interval. For a short interval of time, h, the risk per unit time is $\pi(h)/h$. We define the **rate**, λ, as the limiting value of $\pi(h)/h$ as h gets very small:

$$\lambda = \pi(h)/h \text{ , where h is small.}$$

In contrast with risk, the rate is independent of the length of time and is an instantaneous measure of the subject's liability to the event. The unit of time is arbitrary and could be per year, per week, per day, etc. but must be specified.

Consider an interval of time t consisting of N very short time intervals each of length h (i.e. $t = Nh$). An individual may experience the event of interest during any one of these short time intervals. In any particular short time interval the probability that they experience the event of interest is λh. Note that this probability is **conditional** on their not having experienced the event of interest before the particular short time period under consideration, or being withdrawn from observation for some other reason (e.g. death from other causes or loss to follow-up). Similarly, the (conditional) probability that they do **not** experience the event during that particular time interval is $1-\lambda h$. The probability that the individual is observed for the full time interval, t, **without** experiencing the event is therefore $(1 - \lambda h)^N$; i.e. risk $= 1 - (1 - \lambda h)^N$.

An important property of e is that when x is a small positive number, $1 + x \approx \exp(x)$ and $1 - x \approx \exp(-x)$. For example $\exp(0.02) = 1.0202$. It follows that

$$(1 - \lambda h)^N \approx \exp(-\lambda Nh) = \exp(-\lambda t)$$

(It can in fact be shown that $(1 - \lambda h)^N$ becomes exactly equal to $\exp(-\lambda t)$ as the interval h becomes infinitesimal.) The probability that the individual does not experience the event is therefore $\exp(-\lambda t)$ and the probability that the individual **does** experience the event is therefore $1 - \exp(-\lambda t)$. In other words

$$\text{Risk} = 1 - \exp(-\text{Rate x Time})$$

To see why D/Y is an estimate of the rate, imagine splitting each individual's observation time into a series of short time intervals of length h. Each has the same common probability of disease, π, given that disease has not occurred before the interval. If the total observation time is Y then there will be Y/h small intervals. The estimated value of π is then the number of intervals in which an event is observed, D, divided by the total number of intervals, Y/h, which equals Dh/Y. Since the rate is the risk per unit time (π/h), the corresponding rate is D/Y.

STATISTICAL METHODS IN EPIDEMIOLOGY

PRACTICAL 1

Use of STATA for calculation of measures of effect.

Objectives

By the end of this practical, students will be able to use the STATA commands that calculate the main measures of disease and effect used in epidemiology (risks, rates, rate ratios, and odds ratios).

Students who are unfamiliar with STATA should work on this practical today and refer to the introductory material in the appendix in their own time.

BEFORE STARTING the practical, create a folder in your homespace (eg h:\sme) and copy across data sets and help files from u:\download\teach\sme to this folder.

Once in STATA change to the directory in which you have created for SME by typing the command:

```
cd h:\sme
```

Open a log for your output (a file which will store all the commands and output from your analysis)

```
log using h:\sme\sess1.log
```

1. Calculation of risks and rates

For this part of the practical we will use some of the data from the Whitehall study (dataset A).

The study reports a 20-year follow-up of civil servants. Subjects were recruited into the study over a period of just over two years. Exposures of interest included whether and how much each person smoked, the grade of work, classified as high (administrative, professional, executive) and low (clerical, other), and cholesterol and systolic blood pressure levels at entry to the study. We will examine this data set in greater detail in subsequent sessions.

```
use whitehal
help whitehal
```

Use commands **browse, desc**, **summ**, **tab** and **graph** to examine the variables and answer the following questions:

a) How many deaths (total from all causes) were there? How many from coronary heart disease?
b) What was the number of individuals in each smoking category?
c) Create a new variable **smok2**, taking the values 0 for never or ex-smoker and 1 for smokers:

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

d) Check the new variable is correct by tabulating it against the old one

We will now examine the *risk* of death according to smoking status. It is easy to derive risks using the **tab** command.

Type: `tab all smok2`

To add column percentages, type: `tab all smok2, col`

The column percentage for `all` = 1 gives the risk of death in each group multiplied by 100.

Another way to derive the risk of death is

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

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).

Finally, because we have coded smok2 as 0/1 we can use a command specifically designed for risks

```
cs all smok2
```

The output gives you a two by two table with the same cells as the tab command above and below that the risk estimates. Then it gives you the risk difference i.e. risk among exposed (those coded 1 at smok2) 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 (smoking) and the percentage of outcomes among the whole population attributable to the exposure, assuming that your sample accurately represents the % smokers in the population.

2. Calculation of rates

To use most of the STATA commands for the analysis of follow-up data it is necessary first to define the dates of entry and exit into the study as well as the outcome (or 'failure') variable. This is done with the command `stset`. This is treated in more detail in session 2 but we give a basic explanation here.

The command takes the general form:

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

↑	↑	↑	↑	↑	↑
date exit	outcome of interest	subject identity number	become at risk	date of entry to study	unit of analysis

"**timeout**" and the words in brackets are replaced by the names of the relevant variables.

- The **timeout** variable defines the date of exit – in the Whitehall data set the variable has the name **timeout**
- The **fail** variable is the reason for exit and is coded 0 or missing (denoted by .) for censored observation times. – in the Whitehall data set we will examine the outcomes all cause mortality using the variables **all**, and deaths from coronary heart disease using **chd**
- 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 (it will not be used today –see below)
- The **timein** variable defines the date of entry into the study (called **timein** in the Whitehall data set)
  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 and vice versa (as in the commands below).
- 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 using the Whitehall data the minimum specification is

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

Enter this command. The output reiterates what you have specified, gives the total numbers of observations and fail events and the total observation time accumulated over all in the study. You will notice that some new variables , e.g. _t, have appeared in your data set. These are used by STATA in calculating rates. For example _t = timeout – timein , i.e. person years observed for an individual.

Once the specification is done it is possible to calculate rates using **strate**

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

```
strate smok2
```

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.

If you want the rate per 1000 person-years type:

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

To obtain all-cause mortality rates for the finer categories of smoking type:

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

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)
```

3. Calculation of rate ratios

Rate ratios are derived using **stmh** (always used after giving the stset command).

To compare the death rate among smokers with the death rate among non-smokers type

```
stmh smok2
```

The output shows the rate ratio for current smokers (code 1) compared to never/ex smokers (code 0), together with confidence limits. Check using the output from the **strate smok2** command that the rate ratio is the rate for smokers divided by the rate for non-smokers.

The **stmh** command is also used to calculate Mantel-Haenszel estimates of the rate ratio controlling for the effect of one or more confounding variables. This will be described in a subsequent session.

If you want to save the variables you have created, save the dataset using a different name (to ensure you keep a clean data set available).

```
save h:\sme\whitehal2
```

4. Commands to derive odds ratios

For this part of the practical we will use data from the case-control study of HIV infection in women in Mwanza, Tanzania. We will examine this data set in greater detail in the sessions on case-control studies.

```
use h:\sme\mwanza, clear
```

The cases were all women found to be HIV positive in a cross-sectional survey of 12 communities in Mwanza, Tanzania. Controls were randomly selected from women who were HIV negative. Risk factors of interest included the educational level of the women, and sexual practices including number of sexual partners.

We will examine whether there is an association between HIV infection and duration of education. Type:

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

to examine differences in education levels in cases and controls. The **chi** option asks for a test of the null hypothesis that HIV infection is not associated with educational level against the alternative hypothesis that there is an association.

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**). Type

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

to obtain a 2x2 table of case-control status by educational status (none versus some).

Because the probabilities of selection of cases and controls differ and the probability of selection of controls is often unknown, we cannot calculate odds of disease directly. However, we can estimate odds ratios. This is discussed more in a later session. **st** commands do not apply for case-control data (or cross-sectional data) and you can use the **mhodds** command directly,

```
mhodds case ed2
```

Check that this command does indeed calculate the odds ratio for the 2x2 table you just produced.

5. Missing values

It is important to take account of missing values when analysing data. Try tabulating **skin** (skin incisions or tattoos) and then try the following commands

```
mhodds case skin

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

The **c(2,1)** option tells STATA to make code 1 the reference group and gives us the relative odds for code 2 compared with code 1. Why is there a difference in the output from these two commands?

Now set the missing value for **skin** to system-missing (.)

```
recode skin 9 =.
```

and again type

```
mhodds case skin
```

At the end of the session close the log file

```
log close
```

Remember to tidy up your log by removing errors and duplications before printing it out.

6. 'Do' files

This section describes a facility in STATA that is very useful and may be particularly useful to you during your assessment. One of the buttons in STATA has what appears to be an envelope with a pen on it (8th button from the left). Clicking on this button opens the do-file editor. This is a text editor which allows you to save Stata commands in a do-file (basically a

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 2

Crude and stratified rate ratios

Aim

In this session we will consider the analysis of cohort studies by computing disease (or mortality) rates and comparing them with rate ratios (RRs). The effect of potential confounding variables will be dealt with via stratification methods.

Objectives

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

(i) estimate crude disease (or mortality) rates for a whole cohort and for subsets.
(ii) compare rates in different subgroups using rate ratios (RRs).
(iii) assess the precision of the estimated rates and RRs in terms of confidence intervals.
(iv) examine the effect of potential confounders using stratification methods.
(v) summarize the stratum-specific results via Mantel-Haenszel RRs when appropriate.

Data from a 10% random sample of the Whitehall Cohort Study will be used for illustration (the data are in the Stata file `whitehal.dta`). This cohort study was set up to examine risk factors for mortality in male civil servants from various departments around Whitehall, London. Information on exposure to selected risk factors were obtained by self-administered questionnaire and a screening examination over the period 1967-69; all participants were flagged at the National Health Service Central Registry to identify mortality and emigration.

1. Introduction

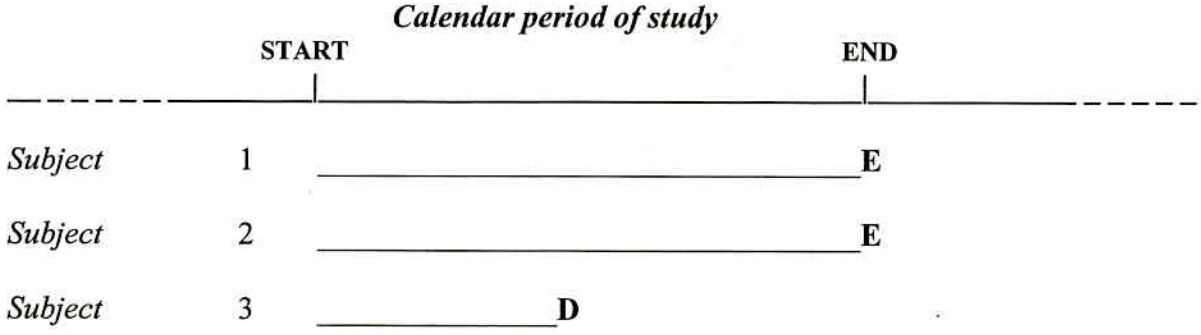
In a cohort study a group of people is followed over a period of time to study the occurrence of an ***outcome*** (e.g. incidence of disease or death). Two types of analyses are generally carried out with this type of study:

(a) when it is possible to categorise the subjects as either *exposed* or *unexposed* to a potential risk factor the rates for the exposed group are compared with the rates for the unexposed group (the topic of this lecture);

(b) when there is no appropriate internal comparison (i.e. a classification into exposed or not exposed is not possible) the rates for the whole cohort are compared with reference rates from a *general population.*

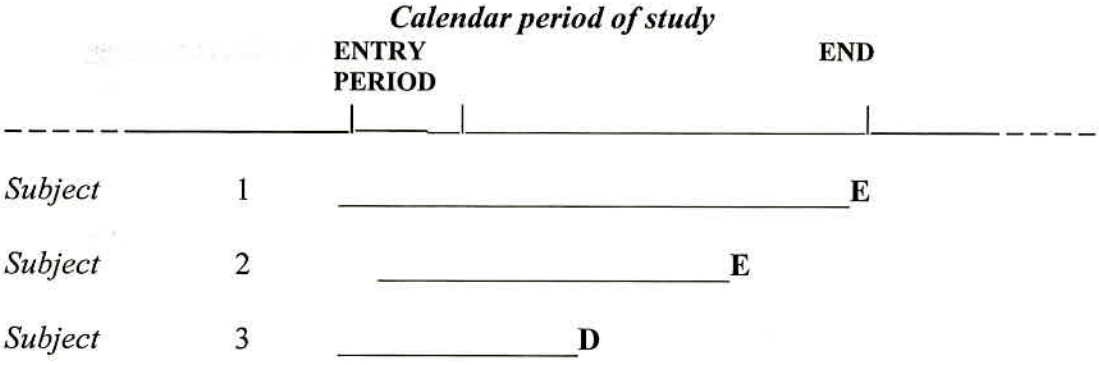
In either case, to calculate incidence rates it is necessary to know the total observation time for all the individuals in the study, i.e. their ***time at risk*** of becoming a case.

The simplest situation is where the entire cohort is enrolled on the same day and all subjects are followed up for a specified time period (that is, there are no losses to follow-up).



Where the symbols are: **E** = exit; **D** = death.

In practice however, individuals may enter or exit at any time during the calendar period over which the study extends and the observation time for each subject is the time between her/his entry and exit. The entry point could be for example: first exposure, disease onset, start of treatment, etc. Exit may be because the subject develops the outcome of interest, or because the subject is lost to follow up, or because the follow-up period ends.



Where the symbols are: **E** = exit; **D** = death.

2. Estimation of the rate for a cohort

To start with we assume that the true rate does not vary during the follow-up period. In this case the rate is estimated by D/Y where D is the total number of events (e.g. death) and Y is the total observation time obtained by adding up the separate observation times for all subjects in the study.

The table below shows death rates from all causes in the Whitehall Study, according to grade of employment (High = administrative/professional; Low = clerical/ other). The values were obtained in Stata with the command **strate grade, per (1000)**, after setting the time and outcome variables using **stset** (see the practical for SME 1).

```
. stset timeout, fail(all) origin(timein) id(id) scale(365.25)
. strate grade, per(1000)

   Grade          D          Y       Rate     95% confidence limits

    High        221    20.3398     10.865        9.523    12.397
    Low         182     7.2656     25.050       21.662    28.966
```

The rates in the high grade civil servants is 10.9 per 1,000 person-years while that for low grade civil servants is 25.1 per 1,000 person-years.

2.1 Confidence limits for a rate

Approximate 95% confidence limits for the true rate, based on D events and Y person years, are found by first computing the 95% error factor EF,

$$EF = \exp\left(1.96\sqrt{(1/D)}\right)$$

where 1.96 is the necessary constant to obtain a 95% interval and $\sqrt{(1/D)}$ is the approximate standard error of the rate (on a log scale; see future lectures on Likelihood).

The lower confidence limit is then obtained by dividing the estimated rate by the error factor, and the upper confidence limit is obtained by multiplying it by the error factor.

Example:
For the High grade workers, the estimated rate is 221/20339.8 = 10.865 per 1000 person-years, the error factor is

$$EF = \exp\left(1.96\sqrt{(1/221)}\right)$$
$$= \mathbf{1.141}$$

and the approximate **95% confidence limits** for the rate are given by 10.865/1.141 = **9.52** and 10.685x1.141 = **12.40**.

Note that if the number of deaths is small (<10) the calculation of the confidence interval should not rely on this approximate method but be based on the Poisson distribution directly (see Likelihood lectures).

2.2 Confidence limits for rate ratio

We can compare the rates estimated for the two grades by taking their ratio with the command:

```
. stmh grade
```

The result is:

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

RR estimate, and lower and upper 95% confidence limits

------------------------------------------------------------
     RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  2.305       73.78        0.0000         1.895       2.805
------------------------------------------------------------
```

The (approximate) 95% confidence limits for this RR are found using a different error factor, which is based on the approximate standard error of the ratio of two rates (on a log-scale),

$$EF = \exp\left(1.96\sqrt{(1/D_1 + 1/D_0)}\right)$$

where D_1 and D_0 are the observed events in the exposed and unexposed groups respectively.

Example:
We compute the error factor,

$$EF = \exp\left(1.96\sqrt{(1/221 + 1/182)}\right)$$
$$= \mathbf{1.217}$$

Approximate **95% confidence limits** for the rate ratio are 2.305/1.217 = **1.895** and 2.305x1.217 = **2.805**.

2.3 Test of null hypothesis

If the rates are the same in the two groups, the true rate ratio is 1. Since the rates are computed as D/Y, we would expect the number of events in each group to be proportional to the total observation time for that group. It follows that the expected number of events in the exposed group, E_1, should be,

$$E_1 = (D * Y_1/Y)$$

where $D = (D_0 + D_1)$ is the total number of events, and $Y = (Y_0 + Y_1)$ is the total follow-up time.

The test is based on the difference between the observed number of events in the exposed group and its expected value, $D_1 - E_1$. The variance of this difference is equal to:

$$V = D * Y_1/Y * (1 - Y_1/Y)$$

Hence the test is calculated as the standardized difference

$$z = U/\sqrt{V}$$

where $U = D_1 - E_1$

and referred to the (standard) normal distribution or, more commonly, as the square of the standardized difference,

$$U^2/V$$

and referred to the $\chi 2$ distribution with 1 df. Note that the letters U and V come from a general notation for significance tests, to be discussed in the Likelihood lectures.

Example (cont'd):
The total number of events was D = 403, of which 182 were exposed (i.e. Low grade);
$E_1 = 403 \times 0.263 = 105.99$, where the value 0.263 is $Y_1/Y = 7265.6 / (7265.6 + 20339.8)$ and
$V = 403 * 0.263 * 0.737 = 78.11$.
The **test** is done by calculating,

$$\mathbf{U} = D_1 - E_1 = 182 - 105.99 = \mathbf{76.01}$$

$$\mathbf{U^2/V} = 76.01^2/78.11 = \mathbf{73.97}$$

This is to be referred to a $\chi 2$ distribution with one df, leading to P<0.0001. (Note that STATA gives a slightly different value: 73.78. This is due to differences in precision.)

3. Exposures with more than two levels

Grade is a categorical variable with two levels. If we wished to look at the effect of a factor, such as systolic blood pressure (SBP), which is quantitative (called **sbp** in **whitehal.dta**), we would need to group its values into, say, 4 groups, giving a categorical variable with 4 levels. For example, we could choose to group the values of systolic blood pressure into <120, 120-139, 140-159, and 160 mmHg and above and call the variable **sbpgrp** (already included in the dataset). These 4 levels are coded 1, 2, 3, and 4 and occur with frequencies:

sbpgrp	level	Freq.	Percent
<120	1	383	22.84
120-139	2	664	39.59
140-159	3	417	24.87
160-	4	213	12.70
Total		1677	100.00

To study how the rate changes with these SBP categories we compute the level-specific rates:

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

sbpgrp	_D	_Y	_Rate	_Lower	_Upper
1	70	6.5188	10.738	8.496	13.573
2	120	11.2343	10.682	8.932	12.774
3	121	6.7018	18.055	15.108	21.576
4	92	3.1505	29.202	23.805	35.822

There appears to be a general increase in rates with increasing levels of systolic blood pressure, as can be seen in the Figure below which was obtained with the command:

```
. strate sbpgrp, per(1000) graph
```

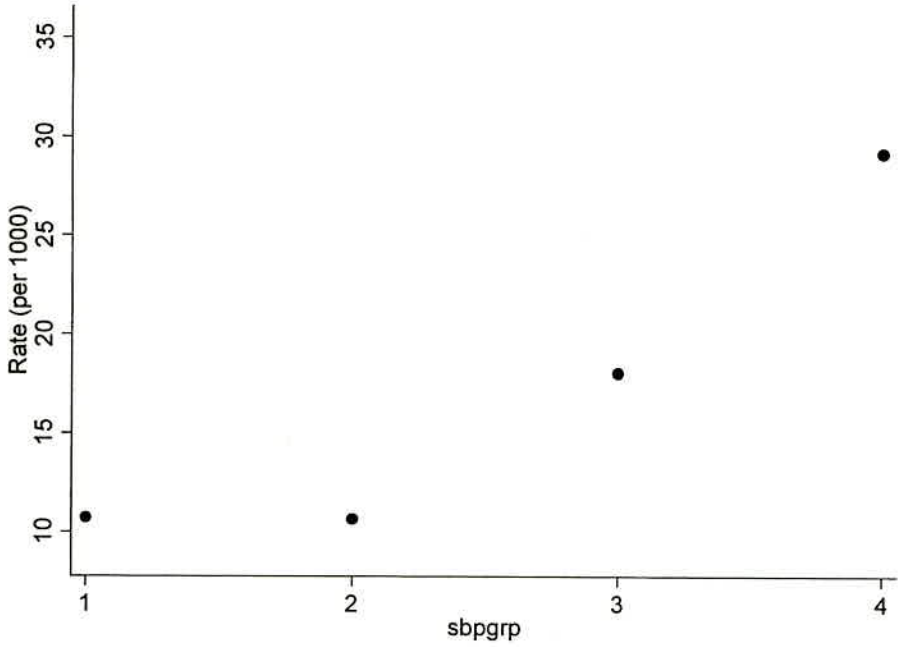


Figure 1. Overall mortality rates by categories of systolic blood pressure.

If we compare level 2 with level 1 we get a rate ratio of 0.995, comparing level 3 with level 2 gives 1.690, and level 4 with level 3 gives 1.617. Each of these rate ratios measures the effect of changing from one level to the next.

GROUP	RR	Lower	Upper	Chisq	p-value
2 v 1	0.995	0.741	1.336	0.001	0.972
3 v 2	1.690	1.313	2.176	16.984	0.000
4 v 3	1.617	1.233	2.121	12.317	0.000

These results are obtained with the commands:

```
. stmh sbpgrp, c(2,1)
. stmh sbpgrp, c(3,2)
. stmh sbpgrp, c(4,3)
```

where, for example, `c(4,3)` stands for "*compare* level 4 with level 3".

The estimated RRs indicate that there is a similar change in rates from level 2 to level 3 and from level 3 to level 4, but not from level 1 to level 2. This can be viewed graphically by using the `yscale(log)` option in `strate`:

```
. strate sbpgrp, per(1000) graph yscale(log)
```

On this scale it is quite clear that there is a linear increase from the second to the fourth group on a logarithmic scale.

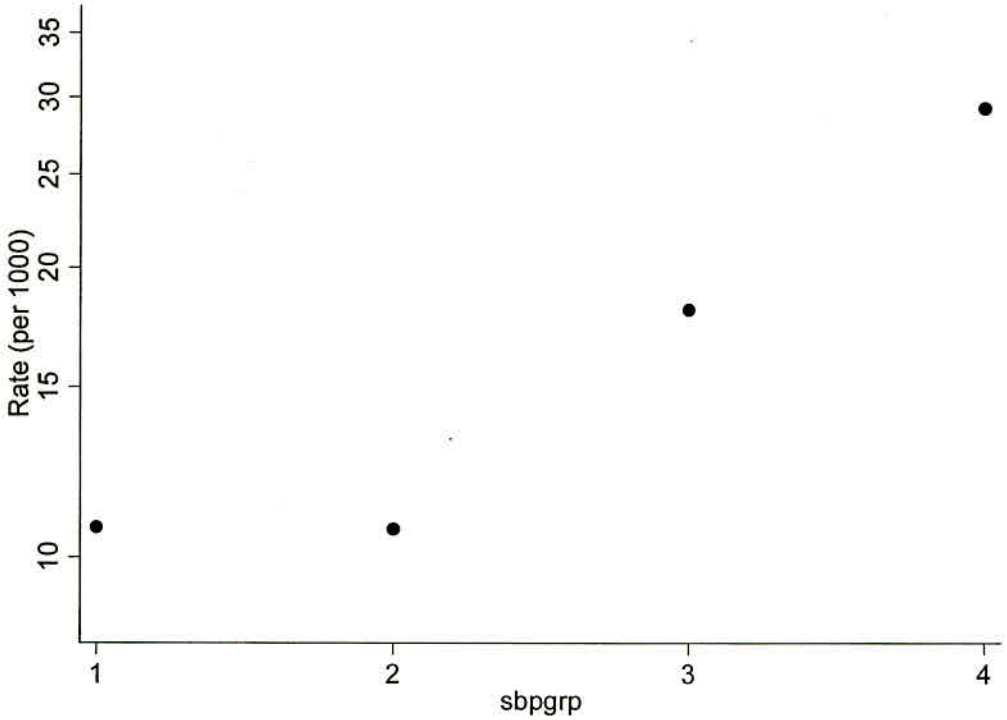


Figure 2. Overall mortality rates (on a log scale) by categories of systolic blood pressure.

If we had forgotten to include the "compare" option when computing the rate ratios, we would have obtained the average effect over these 3 changes in levels (i.e. from 1 to 2, from 2 to 3 and from 3 to 4).

```
. stmh sbpgrp

Score test for trend of rates with: sbpgrp
  with an approximate estimate of the
  rate ratio for one unit increase in sbpgrp

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
1.485	55.52	0.0000	1.339	1.648

The value of 1.485 is the average of the RRs above and represents the common effect across the four categories of systolic blood pressure. In this particular example such summary of the level effects is not appropriate, since we know that there is hardly a change from level 1 to level 2. However, in other cases, it may be a useful summary of the level effects.

4 Stratified analysis of rate ratios

Confounding occurs when there are important differences between the groups being compared which are also related to the outcome of interest. Often there are a number of potential confounding variables in an epidemiological study. These should be investigated to see whether they alter the estimated effects of interest.

The simplest way to examine, and then to control, for confounding is to:

1. stratify the data according to categories of the potential confounding variable,

2. calculate RRs for the exposure of interest in each stratum defined by the categories of the potential confounder,

3. examine whether the RRs for the exposure are similar across the strata defined by the potential confounder or whether they systematically differ,

4a. ***if*** they are similar:
- combine the stratum-specific RRs into a single estimate of the common RR (or RRs, if the exposure has several categories),
- compare the adjusted and the crude RRs, to assess whether confounding was present.

4b. ***if*** they are ***not*** similar:
report separately the RRs estimated for the different strata of the confounder.

Example 1
In the Whitehall study we might be concerned that the effect of grade on mortality is distorted by the confounding effect of smoking practice. Smoking is known to be an important risk factor for mortality, and the prevalence of smoking (at entry to the study) differed between employment grades. For example, more men in low employment grades were current smokers (57%) than were men in high grades (36%).

To adjust the effect of grade for smoking, we must first stratify the data by smoking and estimate the effect of grade separately in each stratum:

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

    smok       RR    Lower    Upper
   never     2.66     1.31     5.41
   ex-sm     1.90     1.33     2.71
   1-14      1.82     1.20     2.77
   15-24     2.44     1.63     3.64
    25+      1.93     1.09     3.41
```

```
Overall estimate controlling for smok
------------------------------------------------------------
     RR       chi2       P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  2.059      52.22       0.0000        1.685       2.515
------------------------------------------------------------

Approx test for unequal RRs (effect modification):     chi2(4) =     1.76
                                                       Pr>chi2 = 0.7791
```

The stratum-specific rate ratios for low grade (coded 2) compared to high grade (coded 1) are shown in the column headed "RR" in the first part of the output; the next two columns show their 95% confidence intervals. The smoking-specific estimates of the RR for grade vary from about 1.8 to 2.7 but do not indicate any systematic trend.

Note:

If there were substantial variation between the stratum-specific rate ratios, the effect of grade would have been modified by smoking. In other words, there would be an *interaction* between the exposure of interest (**grade**) and the stratification factor (**smok**).

The very last part of the Stata output provides a "test for unequal RRs (effect modification)" which can be used to evaluate whether the observed variation between the stratum-specific RRs may have occurred simply by chance. In general, this test is not very powerful, and more understanding is gained by visual inspection of the RRs for the size and pattern of any variation. In the example the test for effect modification is $\chi 5 = 1.76$ (4df), not statistically significant (P=0.779).

So, there is no clear pattern in the stratum-specific RRs and the test is not significant. The assumption of a common rate ratio seems appropriate. Hence we can summarise the 5 smoking-specific estimates of the RR for grade using the Mantel-Haenszel method. This method combines the separate RRs into a weighted average where the weights are determined by the precision with which each RR is estimated (details in the Appendix). The Mantel-Haenszel average of the smoking-specific RRs is shown in the middle of the Stata output and corresponds to the estimate of the RR for grade controlled for smoking. The relevant part of the output is reproduced again below:

```
Overall estimate controlling for smok
------------------------------------------------------------
     RR       chi2       P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  2.059      52.22       0.0000        1.685       2.515
------------------------------------------------------------
```

Note that after controlling for smoking, the rate ratio for "Low" versus "High" grade of employment is reduced from 2.31 to 2.06. This reflects the confounding effect of smoking, that is it indicates that part of the crude estimate of the effect of grade was in fact due to its association with smoking. Note, however, that the smoking-adjusted effect of grade is still highly significant (P<0.001).

There is no statistical test for confounding: confounding is an epidemiological concept that requires the potential confounder to be associated with the exposure of interest and be causally related to the outcome.

Example 2

Another potential confounder for the effect of **grade** is age. The Whitehall data hold a variable called **agein**. We can categorise it into 6 groups and compute the age-specific RRs for **grade**:

```
. egen agecat=cut(agein), at(40,45,50,55,60,65,70)
. stmh grade, by(agecat)
RR estimate, and lower and upper 95% confidence limits

   agecat       RR    Lower    Upper
       40     1.22     0.42     3.57
       45     1.36     0.67     2.75
       50     1.92     1.23     3.01
       55     1.43     1.00     2.06
       60     1.21     0.82     1.80
       65     1.40     0.54     3.62

Overall estimate controlling for agecat

        ------------------------------------------------------------
             RR        chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          1.429       11.36        0.0008         1.160       1.761
        ------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(5) =     2.44
                                                   Pr>chi2 =   0.7854
```

The RRs for **grade** estimated in the six strata defined by age at entry appear to be similar and the test for effect modification is not significant (P=0.79). The Mantel-Haenszel estimate is 1.43 (1.16,1.76) showing that a considerable amount of confounding affected the original estimate of 2.31 (1.90,2.81).

APPENDIX

Technical details about the Mantel-Haenszel method

The Mantel-Haenszel method for combining the separate rate ratios uses a weighted average of the separate rate ratios for the strata. The weights are chosen to give low weight to strata where the rate ratios are poorly determined. The calculations are simple enough to carry out using a hand calculator. Although there is rarely much need to do this these days, it is worth seeing how it is done. We shall illustrate the calculation of the effect of grade controlled for systolic blood pressure.

The rate ratio for any particular stratum is

$$(D_1/Y_1) \div (D_0/Y_0) = (D_1Y_0)/(D_0Y_1)$$

where 1 refers to the exposed group and 0 to the unexposed group. The top and bottom of this fraction can be divided by $Y = Y_0 + Y_1$, the total observation time in the stratum, without altering its value, giving

Rate ratio = RR = $(D_1Y_0/Y) \div (D_0Y_1/Y)$.

The top and bottom of this fraction are referred to as Q and R for short.
So,
RR = Q/R

The first 9 columns of the table below show the data on failures and person-years, together with the quantities Q and R, and their ratio RR.

smok	D0	Y0	D1	Y1	Y	Q	R	RR	U	V
1	21	4603.8	12	988.2	5592.0	9.879	3.711	2.66	6.168	4.801
2	89	8462.4	46	2308.0	10770.5	36.142	19.072	1.89	17.070	22.730
3	43	3161.3	46	1853.4	5014.8	28.998	15.892	1.82	13.105	20.736
4	41	2731.6	57	1559.4	4291.0	36.285	14.899	2.43	21.385	22.671
5	27	1380.4	21	556.3	1936.8	14.967	7.756	1.92	7.211	9.827
	221	20339.8	182	7265.5	27605.4	126.273	61.332		64.941	80.767

The rate ratio for each stratum is equal to Q/R for that stratum. The Mantel-Haenszel combined estimate is obtained from the column totals as $\Sigma Q/\Sigma R$, in this case 126.27/61.33=2.06.

The last two columns are used to test whether the combined estimate of the rate ratio differs significantly from 1, or to find the confidence limits for the combined estimate. The values of U and V in each stratum are the same as those suggested in section 4 to test whether the rate ratio in each stratum differs from 1. Assuming that the true rate ratio is the same in each stratum, the test that its common value is 1 is based on the sum of the Us and the sum of the Vs, in the same way as for each stratum. In this case the chi-squared statistic is $64.94^2/80.77 = 52.2$ ($P<0.001$).

The confidence limits for the Mantel-Haenszel estimate of the common rate ratio are obtained from the error factor

$$EF = \exp(1.96\sqrt{[V/(QR)]})$$

In this case the error factor is

$$EF = \exp(1.96\sqrt{(80.77/(126.27 \times 61.33))} = 1.22$$

so the confidence limits are from $2.06/1.22=1.69$, to $2.06 \times 1.22=2.52$.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 2

COMPUTER PRACTICAL

Aim

In this session we will compute and compare mortality rates with data from a 10% random sample of the Whitehall Cohort Study (`whitehal.dta`) using Stata. We will also learn how to compute Mantel-Haenzsel summary RRs and discuss whether this is an appropriate summary.

Objectives

Having worked through this practical students will be able to:

(i) define the outcome and time variables of the study using `stset;`
(ii) estimate crude mortality rates for the whole cohort and for subsets defined by age at entry using `strate`;
(iii) compare these rates using `stmh`;
(iv) examine whether age at entry confounds the effect of grade.

Exercises

1. Open a log file, change directory to where your SME datasets have been copied and then read the Whitehall data, `whitehal.dta`.

2. In order to analyse overall mortality, set the time and the all mortality outcome variables with `stset` (remember that the time variables are `timein` and `timeout`, the outcome is `all`, the identifier is `id` and the scale should be set to be in years).

3. Investigate how overall mortality varies according to age at entry to the study (you will need to recode `agein` into suitable groups to do this) using `strate`. Use your calculator to verify the confidence interval for the rate in one of the age-groups.

4. Use again `strate` but this time produce a graph showing the mortality trend with age at entry by using the option `graph`. Try also the `yscale(log)` option to plot the rates on a log scale.

5. Use `stmh` to compute rate ratios for the effect of age at entry using the youngest age group as baseline (remember to use the option `c(...)` to compare different categories).

6. Examine the overall mortality rates for low and high grades of employment as given in the lecture notes. Use the `stmh` command to estimate the rate ratio for low grade employees (coded 2) versus high grade employees (coded 1).

7. Use **stmh** to examine the effect of **grade** stratified by age at entry :

```
stmh grade,  by(agecat)
```

8. Is there any evidence of interaction between employment grade and age at entry? Examine the result of the test for interaction.

9. Is the effect of **grade** confounded by age at entry?

10. Examine the effect of employment grade on CHD mortality. To do this you need to redefine the outcome variable (from **all** to **chd**) using **stset**. Is there any evidence of interaction between grade and smoking? Is this effect confounded by smoking?

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 3

INTRODUCTION TO SURVIVAL ANALYSIS

Aim

In this session we shall introduce an alternative approach to the analysis of cohort data. We will be concerned with studying **time to** an event instead of the **rate** with which an event occurs in the population.

Objectives

By the end of the session students will be able to:

(i) To estimate survival probabilities using the Lifetable and Kaplan-Meier methods;
(ii) To compare these two approaches;
(iii) To apply the logrank test for the equality of two survival curves;
(iv) To compare survival analysis methods and classical epidemiological estimates of rates.

Data from a cohort study of men carried out in Trinidad will be used for illustration (the data are in the Stata file `trinmlsh.dta`).

1. Introduction

Most epidemiological cohort studies look at the incidence of relatively rare events/diseases, where incidence rate could be assumed not to vary or to vary only gradually with time (or age). In such cases analyses are usually based on comparisons of rates. However these methods are less appropriate where:

a. There is an interest in quantifying the time elapsed from entry to the event.
Example: time to remission of disease following treatment (median time to remission is of most interest).

b. Incidence rates vary rapidly over time (or age).
Example: time to death following surgery for cancer.

This lecture introduces methods for the analysis of data when either or both these conditions occur.

When time to the event is the main outcome of interest, the relevant data are known as survival data or failure time data (the terminology arises from the statistical methods developed for the analysis of cancer trials and for quality testing in manufacturing). Despite the name, however,

the event under study is not necessarily a failure: it could well be a positive event, like pregnancy.

The survival time for each individual is the time from a predetermined start point, for example entry into the study, until the occurrence of the event of interest. Note that in a clinical trial we could calculate survival time either from the onset of disease, or from the start of treatment. In an observational study we may prefer to calculate survival time from entry to the study or from a fixed age point or from a particular date, for example, the date when first exposed to a carcinogen.

2. Censored Observations

An important feature of data collected in longitudinal studies (either clinical trials or observational) is that the time to the event of interest may be **censored**, that is for some subjects the follow-up may not be complete and the event is not observed to happen.

For example, in the Trinidad study that we will use for illustration in this lecture, subjects were first identified via a cross-sectional survey and then followed up for ten years. During that time dates and causes of death of those who died were recorded. The main event being investigated was death from cardiovascular disease (CVD) and the time until the event occurred was the measure of interest. However the event was not recorded for some subjects when:

- they were still alive at end of follow-up;
- they were lost to follow-up after a certain date (e.g., they had migrated) and hence their vital status is unknown after that date;
- they had died from some other cause. Competitive risk

In the three cases above, the actual survival time (i.e., time to CVD death) is not known. It is known only that the individuals concerned survived until the end of follow-up, or the time of migration, or death from a cause other than CVD. It would **not** be appropriate to exclude such individuals from the analyses, since the fact that they did not die of CVD whilst they were in the study provides some information about survival.

3. Estimation of survival curves

One of the main objectives in survival analysis is to obtain an estimate of the survival experience of the population. This can be achieved by calculating the cumulative survival experienced by the subjects included in a cohort. Two methods, called the Lifetable method and the Kaplan-Meier method, can be used. They are closely related and are illustrated below using probability tree diagrams.

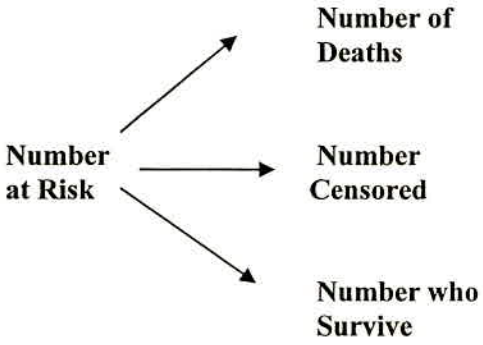
3.1 The Lifetable Method

As an example, we shall use the Trinidad dataset consisting of the 318 men who were aged 60 or over at the time of the survey. Precise date of death was recorded for each subject who died during the follow-up, and all other observations were censored at the end of the follow-up period. There were 88 deaths recorded in this group during the period of follow-up, of which 22 were attributed to cardiovascular disease.
Of the 318 men, 109 described themselves as 'current smokers' while 208 as non- or ex- smokers (the smoking information was missing for one person). We will firstly concentrate on the 109 current smokers for whom the number of deaths from any causes per year of follow-up is summarized below:

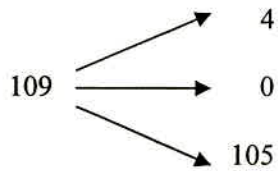
current smokers
(N=109)

Time Band (yrs)	**Deaths from any cause**	**Censored observations**
(0,1]	4	0
(1,2]	5	0
(2,3]	4	0
(3,4]	5	0
(4,5]	7	1
(5,6]	2	4
(6,7]	2	5
(7,8]	7	12
(8,9]	3	27
(9,+)	1	20
All	40	69

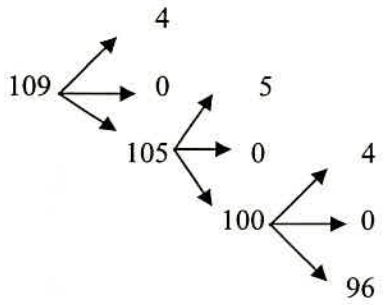
Using a tree diagram, the experience of the cohort in each interval can be illustrated as



For example, *current smokers* in the interval (0,1] year split as follows,



If we extend our diagram to the experience of the first three years we have the following:

	Time band		
	(0,1]	**(1,2]**	**(2,3]**
			
Interval	**1**	**2**	**3**

Note that in each consecutive year the number at risk is computed according to the number of deceased and censored subjects in the previous year. For example, the number at risk at the end of the first interval is 105 = (109-4-0), where zero is the number of censored subjects.

A sensible estimate of the probability of death in the first year is 4/109 and of the probability of surviving the first year is (1 – 4/109). The probabilities of death and surviving associated to each of these three intervals, which we index with the letter ***i***, then are,

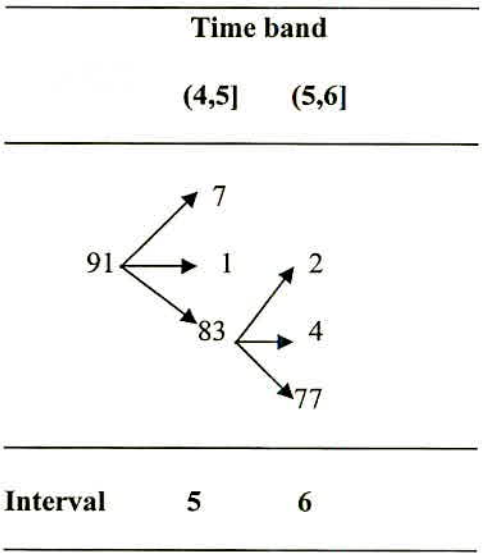
Interval *i*	Time band	Prob(death|interval)	Prob(surviving| interval)
1	(0,1]	4/109= 0.037	0.963
2	(1,2]	5/105=0.048	0.952
3	(2,3]	4/100= 0.040	0.960

Here we have calculated the three separate probabilities of dying in each interval ***i*** as: (4/109)= 0.037; (5/105)= 0.048 and (4/100)= 0.040. The three probabilities of separately surviving each interval then are: (1-0.037)= 0.963; (1-0.048)= 0.952 and (1-0.040)= 0.960. These are called **conditional probabilities**, because in order to be valid, they require that a subject has survived at least up to the beginning of that interval.

To estimate the cumulative probability of surviving to the end of an ***interval i*** we need to multiply the sequence of the year-specific conditional survival probabilities. This is because in order to have survived to the end of any given year a person must have survived all previous years. The cumulative survival to the end of interval 3 is therefore,

$$S(3) = 0.963 \times 0.952 \times 0.960 = 0.88$$

So far we have not had to allow for any censored observations in these calculations. However, the tree diagram for intervals 5 and 6 is as follows:



Note that the subjects who are censored during an interval are not at risk for the whole duration of that interval. We do not know when they were censored but we could assume that, on average, they were lost in the middle of the interval. If we are prepared to assume this, then the

denominator in the formula for the conditional probability of death becomes:

$$Prob(death|interval) = \frac{(Number\ of\ deaths\ in\ interval)}{(Number\ at\ Risk\ -\ (0.5 \times Number\ Censored))}$$

So, in **interval 5**, of the initial 91 subject at risk, 90 people were at risk for the whole year and 1 person for 0.5 year each, on average. Therefore,

$$\text{Prob(death| interval 5)} = 7/(91 - 0.5) = 0.077$$

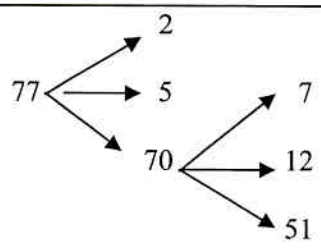
In **interval 6**, of the initial 83, 79 people were at risk for the whole year and 4 people for 0.5 year. Therefore,

$$\text{Prob(death| interval 5)} = 2/(83 - 2) = 0.025$$

The calculation of the interval-specific survival probability and of the cumulative survival, S(i), follows as before. The full listing of probabilities then is:

Interval *i*	Time band	Prob(death|i)	Prob(surv| i)	S(i)
1	(0,1]	0.037	0.963	0.963
2	(1,2]	0.048	0.952	0.917
3	(2,3]	0.040	0.960	0.880
4	(3,4]	0.052	0.948	0.834
5	(4,5]	0.077	0.923	0.770
6	(5,6]	0.025	0.975	0.751

Exercise 1: Calculate the required probabilities for interval 7 and 8 using the same method as above. The probability tree is:



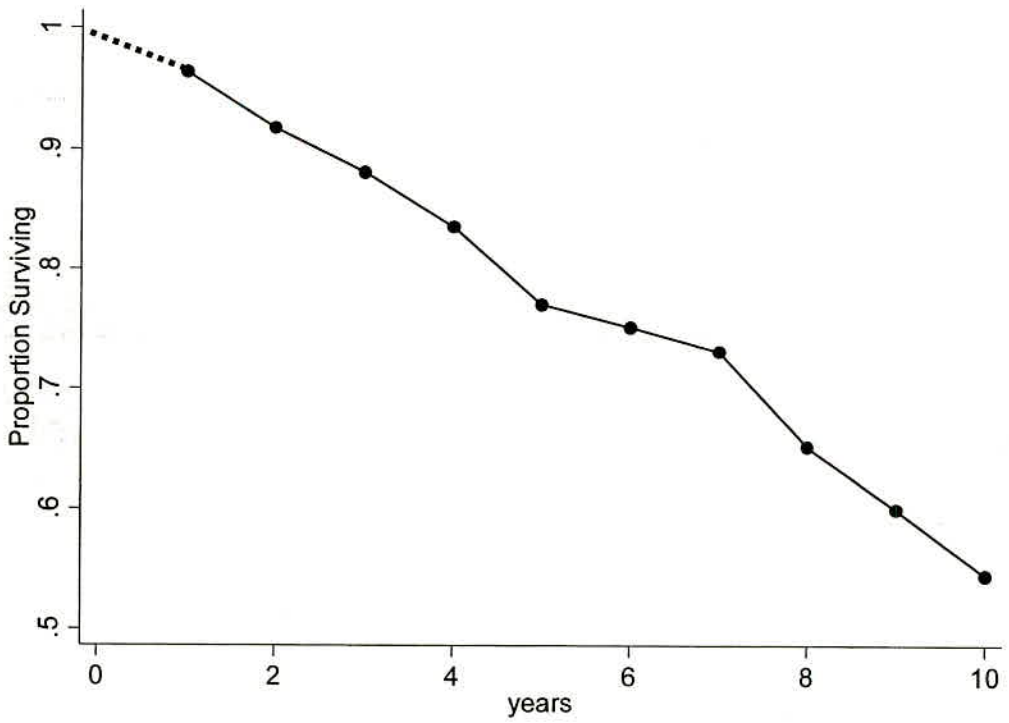
Interval	7	8

From these you compute the conditional and cumulative survival probabilities.

Interval *i*	Time band	Prob(death|i)	Prob(surv| i)	S(i)
7	(6,7]			
8	(7,8]			

This method of deriving the cumulative survival, applied to data grouped in broad time intervals, is known as the **Lifetable method**. The Lifetable itself, as used extensively in the field of demography, is a way of displaying these probabilities in tabular format. Figure 1 shows its content in a graphical form.

Figure 1: Lifetable estimate of the survival curve of current smokers



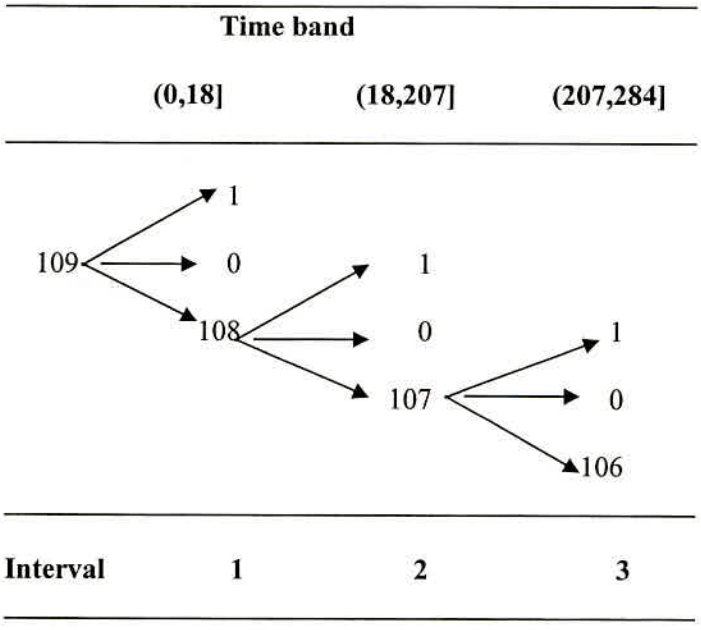
3.2 The Kaplan-Meier Method

In most cohort studies we have precise information on the time of death or censoring of every individual. Hence it would seem preferable to use the whole of this information when calculating survival curves. This can be achieved with the Kaplan-Meier method which is based on a natural extension of the ideas used in the previous calculations.

The following table shows the precise time in years of the first three events.

Time in days	Time in years	Deaths from any Causes	Censored observations
18	0.049	1	0
207	0.567	1	0
284	0.778	1	0

Now that our time interval is small enough to distinguish each individual event, we can concentrate on just the intervals when an event occurs. The tree diagram of the first three events would then be as follows:

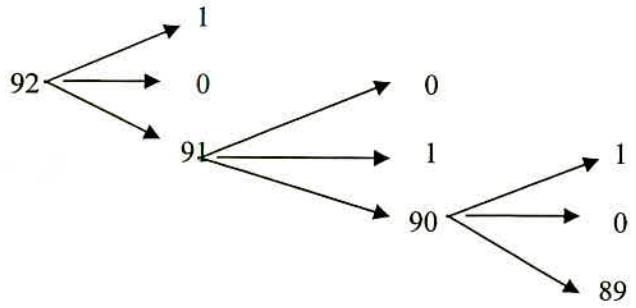


Clearly there are many more calculations to make on the whole set of current smokers, but if we calculate the cumulative survival probability S(i) in the same way as before we will have a more accurate empirical description of the cumulative survival of the cohort.

Exercise 2: Calculate the required probabilities for these three time points using the Kaplan-Meier method.

Interval *i*	Time band	Prob(death|i)	Prob(surv| i)	S(i)
1	(0, 18]			
2	(18, 207]			
3	(207, 284]			

To see what will happen to censored observations we need to jump to the 19th interval in this cohort:

	Time band		
	(1418,1430]	(1430,1526]	(1526,1546]
			
Interval	**19**	**20**	**21**

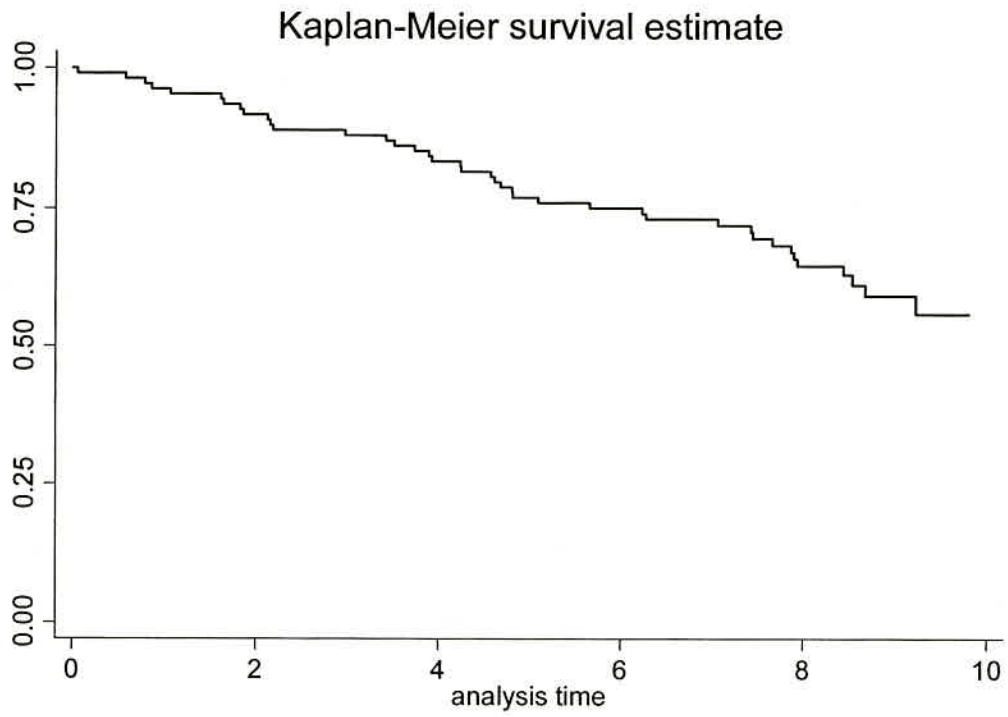
The conditional and cumulative survival probabilities for these intervals are:

Interval *i*	Time band	Prob(death|i)	Prob(surv| i)	S(i)
19	(1418,1430]	0.011	0.989	0.835
20	(1430,1526]	0.0	1.0	0.835
21	(1526,1546]	0.011	0.989	0.826

As we would expect, censored observations do not cause a reduction in the cumulative survival, but adjust the number at risk for the next death to occur. We therefore make no assumption about the time of censoring, but adjust the number at risk at the precise time that the censored observation occurred.

After completing these calculations we would achieve a complete set of cumulative survival probabilities for the current smokers. This information is presented as a survival curve: the plot of S(i) against time is shown in Figure 2.

Figure 2: Kaplan-Meier estimate of the survival curve of current smokers

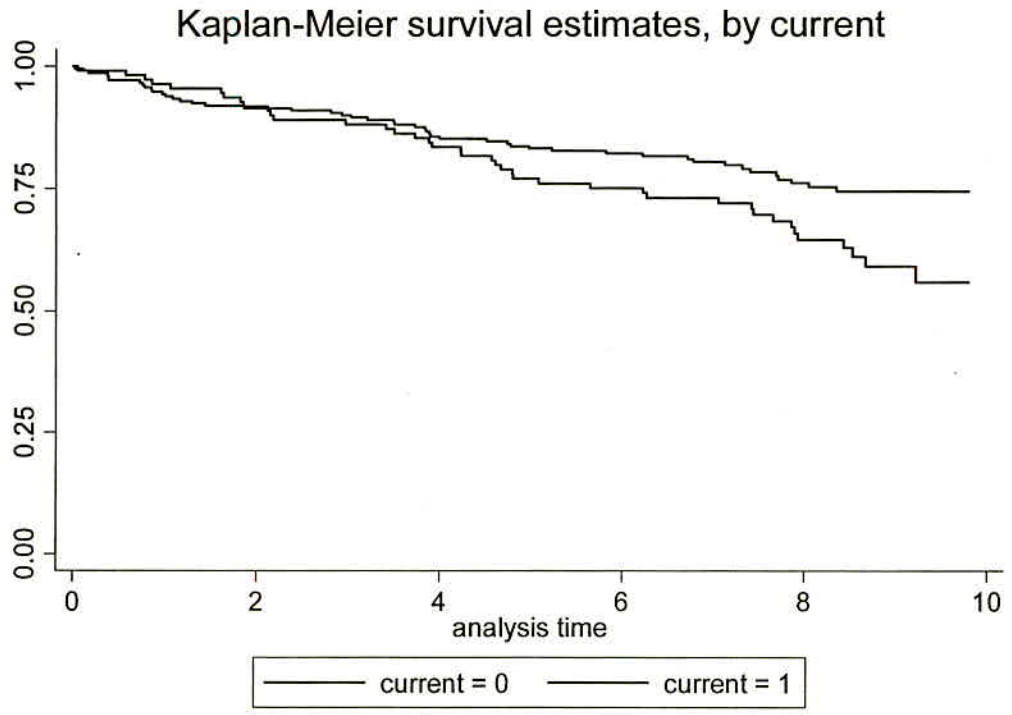


The Kaplan-Meier survival estimate takes the form of a ***step plot*** *rather than a smooth curve, reflecting the fact that this is an empirical estimate of the survival experience of the population at every point in time. By contrast, the corresponding plot of the Lifetable estimates joined the points by straight lines, reflecting the assumption of smooth decline in survival between intervals.*

4. Comparing two groups - The Logrank test

Suppose that we wish to test whether there is any real difference between the survival experience of current smokers and non-current smokers. First we would compare cumulative survival curves for the two groups, using the Kaplan-Meier method since this uses all the available information.

Figure 2: Kaplan-Meier estimate of the survival curve of current and non-current smokers



Then we could test whether they significantly differ from each other using the Logrank test. This method is based on a very simple idea. Let n_{0i} denote the number of subjects at risk at the beginning of the i-th interval and d_{0i} the number of deaths occurring during the i-th interval (usually equal to 1) in the non-smoking group (the unexposed), and n_{1i} and d_{1i} the corresponding values for the current smokers (the exposed). So for the i^{th} interval we write:

	Exposed	**Unexposed**	**Total**
Deaths	d_{1i}	d_{0i}	d_i
Survivors	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
Total at risk	n_{1i}	n_{0i}	n_i

For example the table for the first interval, (0, 18] days, is:

	Exposed	Unexposed	Total
Deaths	1	0 1	
Survivors	108	208	316
Total	109	208	317

For each interval we can then calculated the number of expected deaths in the exposed and unexposed group with the usual formula for 2x2 tables. Summing these "table-specific" expected values and comparing them with the total observed we carry out the test. If the observed numbers differ only by chance from those expected then the survival curves in two groups differ only because of random variation.

In this example we have:

	Observed (O)	Expected (E)
Non-current smokers	48	57.99
Current smokers	40	30.01
Total	88	88.00

The formula for the Logrank test is:

$$\frac{(O_0 - E_0)^2}{E_0} + \frac{(O_1 - E_1)^2}{E_1}$$

Where O_0 and E_0 are the observed and expected numbers of events in the unexposed group and O_1 and E_1 the observed and expected number of events in the exposed group. In this example the test is equal to $(48\text{-}57.99)^2/57.99 + (40\text{-}30.01)^2/30.01 = 5.05$, to be compared with a Π^2 distribution with 1 degree of freedom. It is statistically significant (P=0.03). Thus there is evidence that the survival experience of current smokers is different from that of non-current smokers.

5. Lifetables with Stata

Lifetables are easily produced with the Stata command **ltable**. For example, to estimate the current smokers' survival probabilities of section 2, we type

```
. use trinmlsh,clear
. gen current = smokenum>=2 & smokenum!=.
. ltable years death if current ==1
```

where the variable `current` defines who was a current smoker (at entry into the study) and who was not, the variable `years` holds the follow-up time and the variable `death` holds the death indicator. If one wished 2-yearly intervals instead of 1-yearly intervals, one should use the interval option, `in()`,

```
. ltable years death if current==1, in(2)
```

To tabulate the lifetables for current and non-current smokers at the same time we use,

```
. ltable years death, by(current)
```

To graph these lifetables we use,

```
. ltable years death, by(current) graph
```

6. Kaplan-Meier and Logrank with Stata

We are now familiar with the `stset` commands which allows us to set the time and outcome variables of a longitudinal study. To plot Kaplan-Meier cumulative survival probabilities and to compute Logrank tests we use a command beginning with `st`, i.e. a command that requires `stset` to be defined before it can be used. The command is `sts`; it needs to be followed by the keyword `graph` to plot Kaplan-Meier curves, by the keyword `list` to list Kaplan-Meier survival probabilities, by the keyword `test` to produce a Logrank test.

The relevant time and outcome variable in the Trinidad dataset are `timein`, `timeout` and `death`:

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

So, to produce the Kaplan-Meier survival for current smokers we need to type:

```
. sts graph if current==1
```

and the two curves for current smokers and for non-current smokers,

```
. sts graph, by(current)
```

To list the sequence of cumulative survival probabilities for current smokers we need to type:

```
. sts list if current==1
```

To test whether the survival probabilities of current and non-current smokers differ from each other we type:

```
. sts test current
```

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 3 PRACTICAL

Aim

In this session we shall practice how to compute Lifetable and Kaplan-Meier survival curves and how to interpret their results. We will also practice how to formally compare the survival curves estimated in two subgroups of the same population using the Logrank test.

Objectives

- To produce Lifetable and Kaplan-Meier survival curves using Stata;
- To compare how the estimates obtained from these two methods perform;
- To formally compare the survival curves estimated for different groups using the Logrank test.

The data collected from a cohort study of men carried out in Trinidad will be used (`trinmlsh.dta`).

1. Open a log file, change directory to where your SME datasets are held and then read the Trinidad data. Use `sum` and `describe` to familiarize yourself with the data. In particular identify the names of the variables that hold the outcomes of interest and the follow-up time (type `help trinmlsh` for more details).

2. Examine the overall survival experience of these men, i.e. analyse the outcome called `death`. First compute the survival curve using the Lifetable method. You will need the command `ltable` which, by default, only produces a table by units of time (if you have used the variable `days` to indicate the follow-up time the table will be very long!). In order to additionally produce the survival plot you will need to use the option `graph`. Further, to make the plot comparable to the one you will produce in Question 3, and answer Question 4, use the options shown below (type it all in one line):

```
ltable years death,graph noconf yscale(range(0 1)) title(Lifetable
                                survival estimate) saving(plot1)
```

   Here `noconf` means "do not include the confidence intervals" and `saving(plot1)` saves the plot in a file called `plot1.gph` (if you use this command more than once remember to use "`saving(plot1,replace)`" after the first time).

3. Now use the Kaplan-Meier method to produce the equivalent survival curve for overall mortality in the complete cohort. Remember to **stset** the data first with:

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

To be able to deal with Question 4, save the Kaplan-Meier plot with the **saving** option:

```
sts graph, saving(plot2)
```

4. How similar/different do they look? To look at them side by side use:

```
graph combine plot1.gph plot2.gph
```

To compare their numerical values use the command:

```
sts list
```

which gives the (very long) listing of the Kaplan-Meier curve.

5. You can add 95% confidence intervals to the Kaplan-Meier curve with the option **gwood** which uses the Greenwood's approximation:

```
sts graph, saving(plot2,replace) gwood
```

Compare this plot with the one produced by the Lifetable method (now you do not need the option **noconf**):

```
ltable years death,graph yscale(range(0 1)) title(Lifetable
                            survival estimate) saving(plot1,replace)

graph combine plot1.gph plot2.gph
```

Are they different?

6. Examine the variable **smokenum**. To identify its numerical values use the option **nolabel** as follows:

```
tab smokenum
tab smokenum,nolabel
```

Now, generate a new variable **current** that identifies the participants who were "current smokers" at entry into the study. Compare the survival curves of "current" and "non-current" smokers using either estimation methods (you will need the option **by(current)** in both cases). Are the survival curves of these 2 groups of subjects different?

7. Test whether the survival curves of these two groups of subjects significantly differ over the whole follow-up period using the Logrank test.

8. Repeat the steps used in Questions 6 and 7 using a new smoking variable called **noexcur**, defined as: 0= non-smokers, 1= ex-smokers, 2= current smokers. To generate it you could use the command (and then check the content):

```
recode smokenum 2/5=2,gen(noexcur)
tab smokenum noexcur
```

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 4

CASE-CONTROL STUDIES

Objectives

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

(i) describe the key features of a case-control study
(ii) perform and interpret a crude analysis of a binary exposure in an unmatched case-control study

1. What is a case-control study?

A cohort study follows at-risk individuals over a period of time and identifies individuals who develop the disease/outcome of interest (cases) during that period. In a case-control study, cases which have already occurred are identified, for example through death certificates, hospital records, laboratory reports, etc. In addition, a comparison group is recruited, usually excluding individuals with the outcome of interest. This approach has three potential advantages:

- it removes the need to follow people over time, waiting to see who falls ill and who doesn't;

- it reduces the number of people without the outcome of interest that need to be studied (especially important with very rare diseases);

- in some circumstances it avoids ethical issues over whether to intervene to try to prevent the outcome of interest.

So a case-control study recruits some people **because they have the disease/outcome of interest** and some people **because they don't** (usually).

2. An example of a case-control study

The example we shall use in this session is that of a case-control study performed in Mwanza, Tanzania, to investigate risk factors for *HIV infection* among females aged 15 years or more. Cases were all females found to be HIV +ve during a cross-sectional survey of 12 communities. Controls were randomly selected from among HIV -ve women. They were not matched in any way to the cases. We shall investigate whether a woman's educational level is associated with her risk of HIV infection.

The simplest way of presenting the results of any case-control study is in the form of a 2x2 table.

	Educational level		
	Some formal education	None/adult education only	Total
Cases	140 (74%)	49	189
Controls	311 (54%)	263	574
			763

It would be **wrong** to say that 140 out of 451 women who went to school were infected and so the prevalence/risk of infection in educated women is 31%. To understand why, ask yourself what would have been observed if, instead of recruiting 574 controls, the investigators had decided to recruit 189 controls (1 per case). The number of cases with education would have remained at 140 but the number of educated controls would have decreased, increasing the proportion of educated women who were cases. What we can say is that a somewhat higher proportion of cases (74%) than controls (54%) had been to school. Another way of putting this is that

"schooling is more common among the infected".

While this statement can be made directly from the table, the sort of statement which we are interested in making is that

"infection is more/less/equally common among women with schooling".

Which of these three options – more/less/equally – is the correct one? The answer is that "schooling is more common among the infected" goes with "infection is more common among women with schooling". This intuitive "flip" in logic is fundamental to the interpretation of a case-control study and it is important that you feel comfortable with it. If you don't feel happy with this, try making up some simple numerical examples to convince yourself that it is true.

Given this flip in logic the next questions that we need to ask are "how much more common is infection among women with schooling?" and "could this observed difference have arisen by chance?"

3. Odds and odds ratios

The *probability* that a case in the population in our example had been to school is estimated by

$$140/189 = 0.74,$$

the number of cases having been to school divided by the total number of cases, and can take any value between 0 and 1.

The *odds* that a case had been to school is estimated by

$$140/49 = \quad 2.86,$$

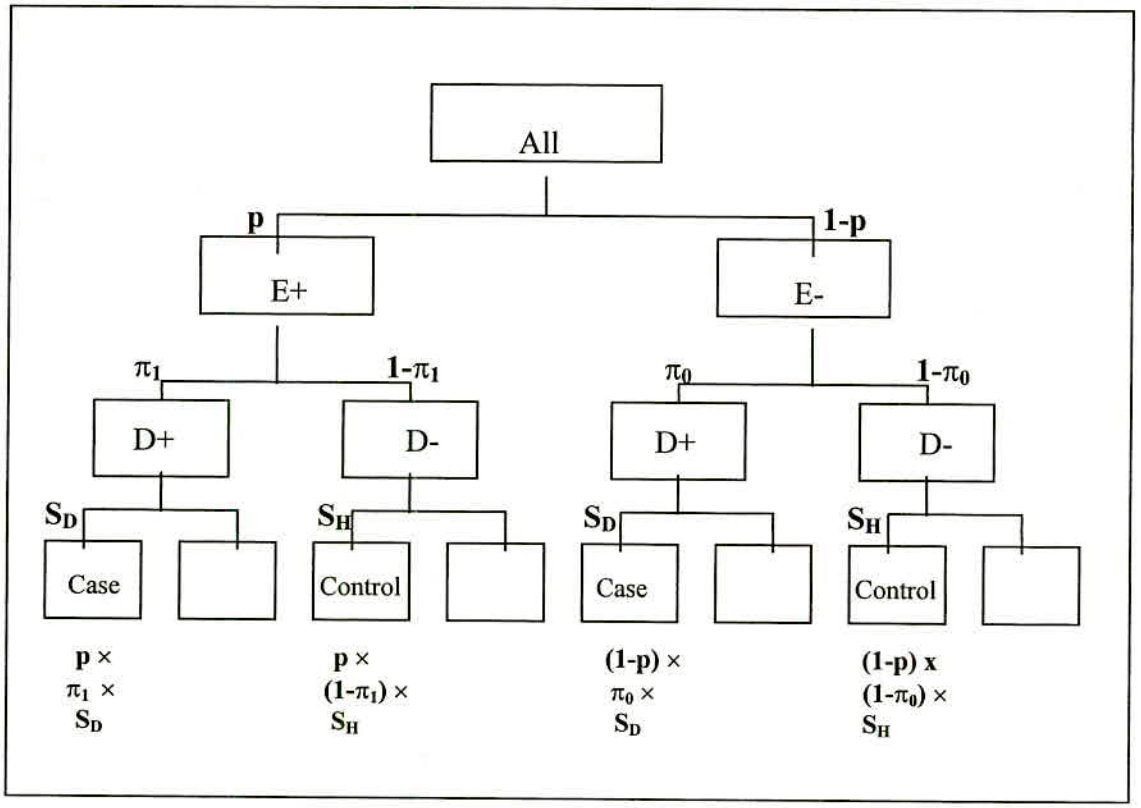
the number of cases having been to school divided by the number of cases not having been to school. Odds (Ω) and probability (p) are related by the formula

$$\Omega = \frac{p}{(1-p)}$$

Ω may take any value between 0 (when p=0) and infinity (when p=1) ($0 \leq \Omega \leq \infty$). We often refer to the odds that a case had been to school as the "odds of exposure among the diseased".

Now consider a population in which a proportion p of the population are exposed (E+) to a risk factor (E). Those exposed have a risk of disease π_1, and those unexposed (E-) a risk π_0. Thus the risk ratio in this population is π_1/π_0.

Suppose that a case-control study is conducted in which a case has probability S_D of being included in the study and a non-case has a probability S_H. In general S_D will be different from S_H (and it is this which makes the study a case-control study). Usually S_D will be much greater than S_H. The "tree" diagram below depicts this situation.



Note that it is assumed that the probability of cases and controls being selected is independent of their exposure status (i.e. S_D and S_H do not depend on whether the individual is exposed or not). This is an extremely important, and usually unverifiable, assumption. Presenting these probabilities in the form of a 2x2 table, we obtain:

	Exposed	Unexposed
Cases	$p \times \pi_1 \times S_D \quad (D_1)$	$(1-p) \times \pi_0 \times S_D \quad (D_0)$
Controls	$p \times (1 - \pi_1) \times S_H \quad (H_1)$	$(1-p) \times (1 - \pi_0) \times S_H \quad (H_0)$

From this table we can derive:

odds of exposure among the cases $= \dfrac{p \times \pi_1 \times S_D}{(1-p) \times \pi_0 \times S_D} = \dfrac{D_1}{D_0}$

odds of exposure among the controls $= \dfrac{p \times (1-\pi_1) \times S_H}{(1-p) \times (1-\pi_0) \times S_H} = \dfrac{H_1}{H_0}$

From these two expressions we can estimate the "exposure odds ratio":

$$= \frac{D_1}{D_0} \div \frac{H_1}{H_0} = \frac{D_1 \times H_0}{D_0 \times H_1}$$

$$= \frac{p \times \pi_1 \times S_D}{(1-p) \times \pi_0 \times S_D} \times \frac{(1-p) \times (1-\pi_0) \times S_H}{p \times (1-\pi_1) \times S_H}$$

$$= \frac{\pi_1 \times (1-\pi_0)}{\pi_0 \times (1-\pi_1)}$$

$$= \frac{\pi_1}{(1-\pi_1)} \div \frac{\pi_0}{(1-\pi_0)}$$

$$= \frac{\text{"odds of disease in exposed"}}{\text{"odds of disease in unexposed"}}$$

$$= \text{"disease odds ratio"}$$

Thus the cross-product ratio from the standard 2x2 table of results for a case-control study estimates the "exposure odds ratio" which is equal to the "disease odds ratio" (providing that S_D and S_H do not depend on whether the individual is exposed or not and cancel out).

What is the "disease odds ratio"? Notice that when the disease of interest is rare, π is small and $\pi/(1-\pi)$ is approximately equal to π (because $(1-\pi)$ is almost 1). So

"disease odds ratio" $= \dfrac{\pi_1}{(1-\pi_1)} \div \dfrac{\pi_0}{(1-\pi_0)}$

$$\approx \pi_1 \div \pi_0$$

which is the risk ratio. So, when the disease under investigation is rare, the cross-product ratio from a case-control study provides an estimate of the risk ratio. Traditionally, this has been the approach to the interpretation of case-control studies and is often referred to as the "rare disease assumption". A more general result states that when the disease is rare (in the

population under study, over the period of the study), the odds ratio, risk ratio and rate ratio are, for all practical purposes, equal (Session 1). It fact, it has been shown that, when the disease is not rare, which of these three measures is estimated by the (cross-product) odds ratio depends on the way in which controls are selected and the rare disease assumption may not be needed (see ASME).

In general terms, we can state that the cross-product ratio obtained from a case-control study provides an estimate of how much more (or less) common the disease/outcome is among the exposed compared with the unexposed.

Returning to our example of HIV infection among women in Mwanza, we estimate the (cross-product) odds ratio as

$$OR = \frac{140x263}{49x311} = 2.42$$

which suggests that the odds of HIV infection are about 2-and-a-half times greater among women who have been to school than among women who have not.

4. Forming a confidence interval for the estimate of the odds ratio

This estimate of the odds ratio (2.42) is subject to sampling variation. To assist us in interpreting this estimate we need to form a confidence interval around it. Calculating confidence intervals "exactly" is complicated and not usually possible on a pocket calculator. A number of computationally simpler approaches have been suggested for calculating an *approximate* confidence interval for the odds ratio (e.g. Woolf's method, Cornfield's method, the test-based method, etc.). When the sample size is "sufficiently large" these methods yield perfectly adequate approximations.

What constitutes "sufficiently large"? There is no established rule for assessing if the data are adequate for the application of approximate methods. As a rough indication, approximate methods should provide reasonable results if each cell of the table (D_1, D_0, H_1, H_0) is greater than or equal to 10. When this condition does not hold many authors suggest that one should present *exact* confidence limits. However, the calculation of these requires a considerable amount of computation which no-one would dream of doing by hand. Some widely used software packages give one the option of calculating exact limits (e.g. EGRET and Statcalc in Epi-Info). Note that STATA *does not* calculate exact limits. Clayton and Hills (Chapter 12) discuss exact methods and draw attention to some of the difficulties associated with their use. They argue that "exact confidence intervals are not exact in any *scientifically useful* sense." Schlesselman (p 180) and Breslow and Day (p 128-129) also provide information on how to calculate exact confidence limits.

One of the simplest approximate methods is that of **Woolf**. This can be derived using the methods described in the sessions on likelihood (see also Clayton and Hills, pp 166-167):

$$\text{Var(logOR) is approximately } 1/D_1 + 1/D_0 + 1/H_1 + 1/H_0 = S^2$$

so that a 95% confidence interval for the odds ratio may be obtained by multiplying and dividing by the error factor

$$\exp(1.96 \times S) \text{ where } S = \sqrt{\{1/D_1 + 1/D_0 + 1/H_1 + 1/H_0\}}$$

Exercise

Calculate a 95% confidence interval around the observed odds ratio of 2.42.

error factor =

The 95% confidence interval for the true odds ratio is then

(2.42 ÷ , 2.42 ×) =

This confidence interval indicates that the data are compatible with a range of situations, schooling being associated with anything from about a 1.7-fold increase in the odds of HIV infection to a 3.5-fold increase.

5. Test of the null hypothesis that the true odds ratio = 1

If schooling is unrelated to the risk of HIV infection, the true odds ratio is exactly 1. To test the null hypothesis that the true odds ratio is 1 (and that the observed odds ratio of 2.42 arose through sampling variation) against the alternative hypothesis that the true odds ratio is not equal to 1, we can apply a chi-squared test for a 2x2 table. As with confidence intervals, the chi-squared test is an approximate test which is perfectly adequate when the sample is "sufficiently large". Because it is approximate there are several different versions (with or without a continuity correction, multiplying by N or by N-1) over which statisticians argue. The version that we present here is that based on the score test (see sessions on likelihood). This is performed by comparing

$$\frac{U^2}{V} \text{ with the } \chi^2 \text{ distribution on 1 degree of freedom}$$

where $U = D_1 - E_1$

and $$V = \frac{D \times H \times N_0 \times N_1}{N^2 \times (N - 1)}$$

[E_1 is the expected number of exposed cases under the null hypotheses;
D is the total number of cases, exposed and unexposed;
H is the total number of controls, exposed and unexposed;
N_0 is the total number of unexposed individuals, cases and controls;
N_1 is the total number of exposed individuals, cases and controls;
N is the total number of individuals in the table]

Using the notation in Kirkwood and Sterne (page 167), U^2/V can be written as:

$$(n-1) \times \frac{(d_1h_0 - d_0h_1)^2}{dhn_1n_0}$$

The only difference between this formula and that presented in Statistics with Computing is the replacement of n by (n-1). This version (which is used by STATA) is known as the Mantel-Haenszel chi-squared test (without continuity correction).

It is an extension of this test which is used to test the null hypothesis that the true odds ratio is 1 when the data are stratified to control a confounding variable (see subsequent session).

Example (continued)

$$U = 140 - \frac{451 \times 189}{763}$$

$$= 28.284$$

$$U^2 = 800.01$$

$$V = \frac{189 \times 574 \times 451 \times 312}{763^2 \times 762}$$

$$= 34.41$$

So $$\chi^2 = \frac{800.01}{34.41}$$

$$= 23.2 \text{ on 1 df}, \quad p < 0.0001$$

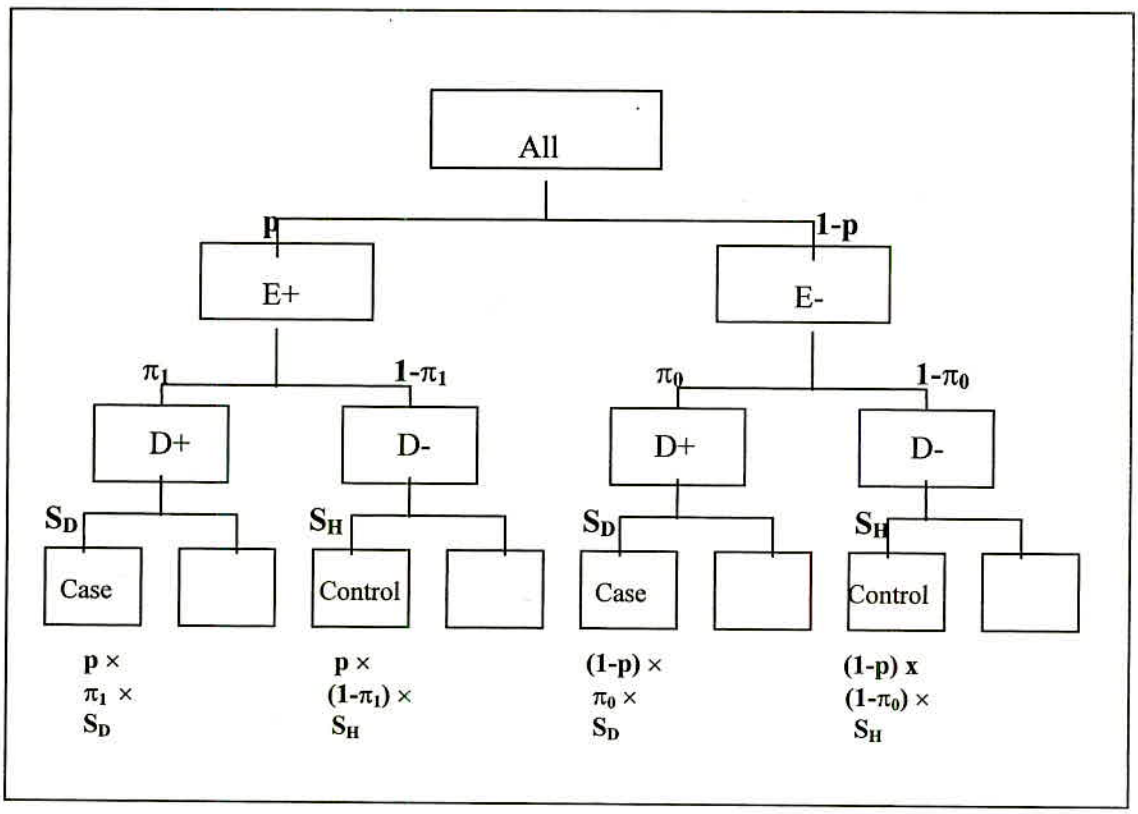
This result indicates that the observed data have very low compatibility with the null hypothesis and would be regarded as strong evidence in favour of an association between schooling and risk of HIV infection in this population. This is consistent with what we observed when examining the 95% confidence interval, namely that the data are incompatible with a true crude odds ratio of 1 (lower limit of C.I. = 1.68). But remember that we have only observed an association between schooling and odds/risk of HIV infection. We would be very unwise (and unepidemiological) to conclude at this stage that going to school increases a woman's risk of HIV (causation).

When the sample size is small (total number of individuals < 20 or total < 40 and one of the *expected* numbers is less than 5) an "exact" test (Fisher's exact test) may be used (see e.g. Kirkwood and Sterne, p 169-171).

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 4 PRACTICAL

Case-Control Studies



1. Consider a population of 100,000 people, 10% of whom are exposed (p=0.10). Let the risk of disease among the exposed be 1% ($\pi_1 = 0.01$) and among the unexposed be 0.5% ($\pi_0 = 0.005$).

a) - what is the true disease risk ratio?

- what is the true disease odds ratio?

b) Now suppose a case-control study is conducted into which all cases are recruited ($S_D = 1$) along with an equal number of controls. Construct a 2x2 table showing the results that you would expect to obtain.

(i) what is S_H?

(ii) what is the estimate of the odds ratio?

c) Now suppose that the probability that a case is recruited is not independent of their exposure status. All exposed cases are identified ($S_{D|E+} = 1.0$) while only 70% of unexposed cases are identified ($S_{D|E-} = 0.7$). Again, 1 control is recruited per case but recruitment of controls is independent of exposure status ($S_{H|E+} = S_{H|E-}$).

(i) what does the 2x2 table look like now?

(ii) what is the estimate of the odds ratio?

(iii) what is this effect called?

2. Now let the risk of disease among the exposed be 10% ($\pi_1 = 0.10$) and among the unexposed be 5% ($\pi_0 = 0.05$).

(i) what is the true disease risk ratio?

(ii) what is the true disease odds ratio?

Explain the difference between your answers to questions 1 and 2.

3. The table below presents data from Mwanza on the association between HIV infection and whether or not the woman has a constant partner/spouse. Analyze and interpret these data.

	Constant partner		
	Yes	No	Total
Cases	131	58	189
Controls	462	111	573

STATISTICAL METHODS IN EPIDEMIOLOGY

Practical 5

Data checking/editing, and univariate analysis

Objectives

By the end of the session students will have:

(i) revised what measures of effect are appropriate for different study designs
(ii) appreciate that classifying variables into exposures/confounders/effect modifiers can help to structure an analysis
(iii) appreciate the importance of data cleaning, and how this should be approached
(iv) be aware of some guidelines for data reduction
(v) appreciate the importance of starting with univariate analyses
(vi) be aware of the dangers of data dredging, data-driven comparisons, and sub-group analyses
(vii) appreciate the value of making analyses reproducible

The practical combines notes on how to approach the initial stages of your analysis, with practice in applying the approach to real data.

The data set dietsme2 will be used throughout the practical. You should begin by going to the introductory part of the manual to read about this data set, and/or looking at the help file dietsme2 (type help dietsme2 in STATA). The Excel data set dietsme2originaldata.xls contains the correct data, and the dataset dietsme2 contains data as entered onto the computer, prior to data checking and cleaning.

The question that you are asked to investigate is:
'Is total energy intake associated with the risk of coronary heart disease?'
The hypothesis is that the risk of heart disease will decrease with increasing energy intake, because energy intake is regarded as a marker for physical activity.

Introduction

Most sessions in this course focus on specific statistical techniques for selected epidemiological situations. In the context of an actual epidemiological study, it can be difficult to decide which procedures to apply and in what order.

Further, while the lecture notes and practical sessions include simple descriptive analyses, the emphasis is on more sophisticated techniques (Mantel-Haenszel methods and regression) that are needed for efficient estimation of the effect of particular exposures, controlled for confounding. These techniques should be used only after data have been cleaned and explored using simple, descriptive approaches such as graphs and cross-tabulations.

This session, together with Session 15, aims to present some general guidelines on how to approach the analysis of data from epidemiological studies, from the decision as to what measure of effect to use, through initial data cleaning and exploration to regression modelling. This session goes as far as univariate analysis of a "cleaned" data set.

1. Measures of effect

For cohort and cross-sectional studies, a decision is needed as to which measure of effect to use. The possible choices for each of the 3 main study designs used in epidemiology are shown in the table below.

Type of study	*Measure of disease frequency*	*Measures of effect*
Cohort (using person-time data)	Rate	Rate ratio Rate difference
Cohort (not using person-time data)	Risk	Risk ratio Risk difference
	Odds	Odds ratio
Cross-sectional	Prevalence	Risk (prevalence) ratio Risk difference
	Odds	Odds ratio
Case-control		Odds ratio

This course concentrates on ratio measures of effect, which are the most commonly used for aetiological enquiries. For most techniques, however, corresponding methods exist for difference measures (see Rothman and Greenland, chapter 15).

For **cohort studies**, the rate is the preferred measure of disease frequency providing person-time data are available, and the rate ratio is generally the preferred measure of effect.

For cohort analysis based on risks, and for **cross-sectional studies**, the risk ratio would generally be regarded as more easily interpretable than the odds ratio. Nevertheless, the odds ratio is often used, because the statistical properties of procedures based on the odds ratio are generally better. Since logistic regression gives estimates of odds ratios, the odds ratio is sometimes also used because it will be consistent with results from logistic regression analysis.

In **case-control studies** the odds ratio is always used as the measure of effect, although this may estimate the risk ratio or rate ratio, depending on the method of selection of controls. If a disease is rare, the odds ratio, risk ratio, and rate ratio are almost identical.

(i) What type of study are the diet data from?
(ii) What is the outcome variable for the study question?
(iii) What measure of effect should you use to analyse these data?

2. Classification of explanatory variables

It is generally helpful to distinguish in advance between 'exposure variables', 'confounders' and 'effect-modifiers'.

Exposure variables are those variables of central interest, whose effect on risk we wish to examine and estimate.

Confounders are variables which distort the relationship between the outcome and one or more of the exposure variables. We collect data on known (and potential) confounders so that we can remove their confounding effect in the analysis.

Effect-modifiers are variables which modify the effect of an exposure variable on the outcome. We collect data on these to examine how the effect of an exposure of interest varies according to the value of the effect-modifier.

In practice, this classification is over-simplified. For example, a variable may confound the effect of one of the main exposures of interest, but its effect may also be of interest in its own right. A variable may be a confounder for one exposure variable, and an effect-modifier for another.

Also, most studies have an exploratory element, in that data are collected on some variables which are not of central interest but which it is thought may turn out to be associated with the outcome. Such variables need to be considered as **potential confounders or effect-modifiers**.

It is also important to think about causal pathways. For an explanatory variable to be considered as a confounder, it must 1) be associated with the exposure under study 2) be associated with the outcome of interest and 3) should not be a mediating factor (intermediate variable) – i.e. it should not be a link in the causal chain leading from the exposure to the outcome.

For example, suppose the outcome is cardiovascular disease, and the exposure of interest is fruit and vegetable intake. If it is thought that part of the effect of fruit and vegetable intake

on cardiovascular disease is through its effect on plasma vitamin C level, it would be incorrect to control for plasma vitamin C level when estimating the effect of fruit and vegetable intake. On the other hand, consider alcohol as an exposure variable for the outcome lung cancer. It is well established that smoking is a risk factor for lung cancer, and smoking and alcohol consumption are often associated. Smoking is not part of any causal pathway linking alcohol to lung cancer, so in this case smoking is a confounder and should be controlled for when investigating the effect of alcohol on lung cancer.

For a clear and detailed explanation of criteria for a variable to be considered a confounder, see Rothman (1986, pp89-94). For a clear, concise discussion of how a 'conceptual framework' can help to structure an analysis, see Victora et al (1997) 'The role of conceptual frameworks in epidemiological analysis: a hierarchical approach', International Journal of Epidemiology, Vol 26 pp224-227.

(i) Given the research question, classify each explanatory variable in the diet data set as one (or more) of the following
a) an exposure (variable of central interest to the research question)
b) a known or potential confounder
c) a known or potential effect-modifier

3. Data editing

Careful checking and editing of the data set are essential before statistical analysis commences. Checks should generally be made separately in individuals with and without disease, as the distributions may be quite different.

A good way to start is to
a) Use the command **describe** to obtain a description of the data set. This will show you the number of observations in the data set, the number of variables in the data set, and give a listing of the variables together with the variable labels. It will also tell you whether a variable is stored as a number, or as text.
b) Use the command **summarize** to obtain summary statistics for each *numeric* variable in the data set. This command gives you the number of observations with data on each variable, and the mean, standard deviation, and range for each variable. If the number of observations with data on a particular variable is less than the number of individuals in the data set, then there are missing data for this variable. If the minimum or maximum values look implausible or impossible, then this suggests there are errors in the data.

(i) Use the commands **describe** and **summarize** to obtain basic information about the diet data set.

The next step is to examine the distribution of each of the variables to check for possible errors.

For categorical variables, this means checking that all observations relate to allowed categories, and that the frequencies in each category make sense. For categorical explanatory variables, it is best to tabulate them separately for diseased and non-diseased individuals (as their distribution may differ between these 2 groups).

(ii) Which variables are categorical in the diet data set?
(iii) Tabulate the outcome variable (chd) to check that all individuals are coded 0 or 1. Tabulate each categorical explanatory variable, separately for those with and without disease. e.g.
Occupation is a categorical variable

```
tab job if chd==0
tab job if chd==1
```

For quantitative variables, range checks should be performed to search for values falling outside the expected range. Histograms or box plots are a good way to look for 'outliers' that look extreme relative to the rest of the data.

(iv) Which explanatory variables are quantitative in the diet data set?

(v) For each quantitative explanatory variable

1) generate a histogram showing the distribution of the variable, separately for individuals with and without disease.
2) List individuals for whom the data could be in error and should be checked – i.e. list individuals for whom the data look extreme relative to the rest of the population.
3) Check the data for these individuals, using the Excel data set dietsme2originaldata.xls – this data set contains the true values for each individual, while dietsme2 contains the data as entered into the computer.
4) Correct the data that are in error.

e.g. height is a quantitative explanatory variable

1) Graph the data

```
histogram height if chd==0, fraction start(70) width(10) xlab(70(10)190)
histogram height if chd==1, fraction start(70) width(10) xlab(70(10)190)
```

(note: see a) Optional Practical 1, Introduction to STATA and/or b) type help histogram in STATA, for more information on the histogram command. xlab covers the range of heights recorded in this data set (see minimum and maximum values of height from the **summarize** command) and labels the x-axis, width(10) tells STATA that the data should be grouped into categories of 10 units, start(70) tells STATA to start dividing the data up from the value 70.

2) List individuals with heights that are extreme relative to the rest of the population – e.g. those with height <=90cm.

```
list id height if height<=90
```

3) Check the data, using the file dietsme2originaldata.xls
For height check the data for individuals 112 and 323.

4) Correct the data in dietsme2 that are in error
For height, the data for individuals 112 and 323 were in error

```
replace height=179 if id==112
replace height=187 if id==323
```

Repeat the above steps for the other quantitative variables.

The second step is to conduct consistency checks, to search for individuals where two or more variables are inconsistent. For example, if sex and parity are recorded, a cross-classification of the two can be used to check that no males were recorded with a parity of one or more. If the study is a cohort one, then the dates of entry and exit should be consistent, and the age at entry should be compatible with any age restrictions on recruitment into the trial.
Scatter-plots can be useful for checking the consistency of quantitative variables; for example, of weight against age, or weight against height. Further outliers can be detected in this way.

(vi) Check that the birth, entry and exit dates are consistent.
List individuals whose date of birth is after the date of entry to the trial

```
list id dob doe dox if dob>=doe
```

List individuals whose date of entry to the trial is after the date of exit

```
list id dob doe dox if doe>=dox
```

Check the data on individuals where the birth, entry and exit dates are not consistent. Correct the values where they are in error, using the **replace** command.
e.g. to correct the date of entry for id 251,

```
replace doe=mdy(2,16,1962) if id==251

(in the mdy command the first number is the month, the second number is the
day, and the third number is the year)
```

(vii) Calculate age at entry into the trial (using the variables dob and doe), and check that all individuals are in the age range 30-67 (study participants were restricted to this age range).

```
gen ageentry=(dox-dob)/365.25
histogram ageentry if chd==0, fraction start(0) width(5) xlab(0(5)70)
histogram ageentry if chd==1, fraction start(0) width(5) xlab(0(5)70)
```

Check the dates of birth and dates of entry for individuals for whom age at entry appears to be in error, using dietsme2originaldata.xls. Correct the data that are in error in dietsme2, using the **replace** command.

Then recalculate age at entry, once you have corrected the dates of birth and dates of entry data.

```
drop ageentry
gen ageentry=(dox-dob)/365.25
```

Check that the data now appear reasonable

```
histogram ageentry
```

(viii) Use scatter plots to conduct consistency checks that you think are relevant
e.g. to check for consistency between height and weight data,

```
scatter weight height if chd==0
scatter weight height if chd==1
```

Some data may still appear extreme after data checking based on looking at the original field forms, since there could be data recording (as opposed to data entry) errors. If it is certain that the data were recorded wrongly (e.g. an impossible weight), then the data should be changed to a missing value code on the database.

In borderline cases, where an observation is an outlier but not considered impossible, it is generally better to leave the data unchanged. Strictly speaking, the analysis should then be checked to ensure that the conclusions are not affected unduly by the extreme values. In practice, quantitative variables are generally grouped into categories before analysis, and one or two outliers are therefore unlikely to have a marked influence on the results.

Once the data have been cleaned as thoroughly as possible, the distributions of each of the variables should be re-examined, firstly to check that all now appears to be in order, but secondly to get a 'feel' for the data — that is, to get a good understanding of the characteristics of the study population with respect to the exposures and other variables measured.

(ix) Recheck the data after you have made corrections, using cross-tabulations, histograms, and scatter plots as above.

(x) Calculate BMI for each individual, as this is a better marker of risk than weight, as it is standardised for height. Note that BMI = weight(kg)/(height(m))^2. So, since height is given in cm in the diet data set, use

```
gen BMI = weight*100*100/(height*height)
```

Check that the distribution of BMI looks reasonable.

4. Data reduction

Before commencing the formal analysis, it may be necessary to group the values of some of the variables. Since the 'classical' methods based on stratification are recommended before moving to regression methods, grouping is essential for quantitative variables. But some grouping may also be necessary for categorical variables with large numbers of categories (e.g. parity, occupation) and/or few individuals/cases in some categories.

An important principle to observe when merging categories is that the risk of disease is expected to be similar within each of the merged groups.

How many groups should be used? This depends partly on the type of variable.

For **exposure variables**, where we wish to examine the pattern of dependence of risk on the degree of exposure, it is an error to use too few categories. The 'unexposed' should generally be treated as a separate category (e.g. non-smokers), and the 'exposed' should be divided into several groups (four or five should usually be enough to give a reasonable picture of the risk relationship).

For continuous variables like blood pressure, one strategy is to divide the range of the variable into, say, quintiles, giving five groups with equal numbers of subjects in each group. This helps to ensure that estimates of effect for each category are reasonably precise, but can

sometimes obscure an important effect if a few subjects with very high levels are grouped in with others with more moderate levels. Alternatively, cut-off points may be chosen on the basis of data from previous studies, the aim being to define categories within which there is thought to be relatively little variation in risk.

Centiles are particularly useful if you have little idea what categories make sense in terms of their being little variation in risk within a category, as they enable you to simply divide the data distribution into equal-sized groups.

If there is a fairly standard way of grouping a variable then try to stick to this. For example, when grouping age it is more natural to use age bands such as 15-19, 20-24, 25-29 rather than base your categorisation on centiles and have a grouping such as 15-18.8, 18.9-23.7, 23.8-29. Also, round numbers where you can e.g. use <10g rather than <9.78g as a lowest category (even if 9.78 corresponds to a percentile of the data distribution).

For variables of interest only as **confounders**, two or three categories may be sufficient to remove most of the confounding. However, more categories will be needed if the confounding is strong, as would often be the case with age, for example. It is often necessary to examine the strength of the association between the potential confounder and the outcome variable before deciding on the number of categories to be used in analysis. The weaker the association, the more one may combine groups.

A further consideration is that a group in which there are no cases/all individuals are cases must be combined with others before inclusion in analysis using Mantel-Haenszel techniques or regression. If it is felt that this group is substantively different from other groups for the risk of disease, however, it may be preferable not to combine them with another group, but instead exclude them from the regression analyses and simply report the total number of individuals in this group and that there were no cases of disease/all individuals had disease. As a general guideline, there should be at least 5 individuals in each cell of the cross-tabulation of an explanatory variable by outcome yes/no.

(i) Categorise the explanatory variables energy, height, BMI, age at entry to the trial. Use histograms of energy, height, BMI, and age to help you group them. You may also find the centile command useful, if you wish to use tertiles, quartiles etc to help you group the quantitative explanatory variables – see help centile in STATA for how to obtain percentiles. Use the recode command to create the new, categorised, variables, and use the table command to check you have recoded the new variable correctly

e.g. for height

```
histogram height, fraction start(150) width(5) xlab(150(10)200)
centile height, c(33 67)
```

* this command gives tertiles of height

```
gen htcat=height
recode htcat min/169=1 170/176=2 177/max=3
```

* this command creates 3 categories of height, 1 "152-169cm" 2 "170-176cm" 3 "177-190cm"

To check your recoding, use

```
table htcat, c(min height max height freq)
```

5. UNIVARIATE ANALYSIS

A) Cross-tabulations
B) Measure of effect

5A) Cross-tabulations

(a) Association between explanatory variable and disease

1) **Case-control studies**: cross-tabulations of each explanatory variable by case/control status should be done, with the **percentage in each category for each of cases/controls** used to summarise the (crude) association between the explanatory variable and disease.
e.g. `tab casecon job, row`
where `casecon` is the variable coding whether an individual is a case or control.

2) **Cross-sectional, and cohort studies/randomised controlled trials** where the outcome measure is a risk: cross-tabulations of each explanatory variable by disease yes/no should be done, with the **percentage who have the disease in each category** used to summarise the (crude) association between the explanatory variable and disease.
e.g `tab job chd, row`
where `chd` is the outcome variable and is coded 1 if yes and 0 if no.

Important: note the difference between 1) and 2) in how the crude association between an explanatory variable and disease should be summarised.

3) **Cohort studies/randomised controlled trials where the outcome is a rate**: the rate of disease should be calculated for each category of each explanatory variable.
In STATA, we need to stset the data before we can calculate a rate (see Session 2).

(i) Use the `stset` command to tell STATA the outcome variable, the date of entry to the cohort, and the date of exit.

```
stset dox, fail(chd) enter(doe) scale(365.25) id(id)
```

(ii) Calculate the rate of disease in each category of each explanatory variable.
e.g. to calculate the rate of disease in each category of height,

```
strate htcat, per(1000)
```

(b) Understanding where confounding may be present
For an explanatory variable to confound the estimated effect of a particular exposure, it must (1) be associated with the outcome and (2) be associated with the exposure (see section 2 on classification of variables).

From part (a) above, you will have an initial idea of which explanatory variables are associated with the outcome. To assess (2), you should cross-tabulate each explanatory variable against each exposure variable.

In case-control studies these cross-tabulations should be done among the controls only.

(i) assess whether height, BMI, age at entry, and job may confound the estimated effect of energy intake on coronary heart disease.

5B) Measure of effect

The next step is to calculate a rate ratio, risk ratio, or odds ratio to summarise the effect of each explanatory variable on the outcome. In STATA, we can use the Mantel-Haenszel commands for this. For a cohort study, we use the command **stmh**, and for a cross-sectional or case-control study we use the command **mhodds**.

To calculate a measure of effect for each explanatory variable, one category has to be chosen as the 'baseline'. Often this will be the 'unexposed' or, if everyone is exposed to some extent, the group with the lowest exposure. If there are very few people in this group, however, this will produce effect estimates with large standard errors, and in this case there can be advantages in choosing a larger group as the baseline. If there is no clear ordering of the categories of the exposure, then it is usual to choose the category with the largest number of individuals as the baseline group.

If the exposure variable is an ordered categorical variable, then a test for trend to examine for evidence of increasing or decreasing risk with increasing levels of exposure may be more appropriate than a test for heterogeneity.

(i) Calculate a measure of effect for each of energy intake, height, BMI, age at entry and job.
e.g. Suppose category 1 of BMI is the baseline category. Then to obtain a rate ratio comparing the rate in category 2 with the rate in category 1, we can use

```
stmh bmicat, c(2,1)
```

Although these crude analyses will be superseded by analyses that take into account the effects of other variables, they are still a very important stage of the analysis because:

1) they give an initial idea of which explanatory variables are strongly related to the disease outcome

2) they given an initial idea of which explanatory variables may confound the effect of each exposure variable

3) the degree to which the crude estimate of effect is altered when other variables are taken into account is a useful indication of the amount of confounding present (or at least, the amount that has been measured and successfully removed).

ADDITIONAL NOTES

1. Difficulties in analysis and interpretation

Consider a large, randomized study of two interventions. The comparison of interest is likely to have been specified in a protocol which was agreed before the study started. If the study is large enough then randomization will ensure that the groups are balanced with respect to all potential confounders. The analysis of such a study is typically straightforward!

A more typical epidemiological study, even where the main exposures are specified in advance, may involve the measurement of many variables. Different choices of variable groupings and modelling strategies may make important differences to the conclusions.

There are contrasting viewpoints on how data from more exploratory studies should be treated. Three reasons for caution are:

(i) Multiple comparisons/data dredging:

Even if no exposure variable is truly associated with the outcome variable, we expect one in twenty independent comparisons to be statistically significant at the 5% level. Thus the interpretation of associations in a study in which the effect of many exposures was measured should be much more cautious than that for a study in which a specific *a priori* hypothesis was specified. Searching for all possible associations with an outcome variable is known as 'data-dredging'.

(ii) Subgroup analyses

We should be cautious about the interpretation of apparent associations between an exposure and disease in subgroups of the data, particularly when there is no statistically significant evidence of an interaction. It is extremely tempting to emphasise an 'interesting' finding in a subgroup in an otherwise negative study. Again, if we look in enough different subgroups, one will emerge with a "statistically significant" association even when there is no association in reality.

(iii) Data-driven comparisons

A related problem is that we should not group the data in order to produce a statistically significant difference, then interpret the p-value as if this had always been the intended comparison. For example, if we have 10 age groups, then we could compare 1 with 2,...,10 or 1 and 2 with 3,...10 and so on. If we choose a particular grouping out of these 9 possible ones because it shows the largest difference between 'younger' and 'older' individuals, then we have chosen the smallest P-value from 9. It is sensible to decide how variables will be grouped as far as possible before seeing how different groupings affect the conclusions of your study.

These problems **do not** mean that all epidemiological studies must have hypotheses and methods of analysis which are specified at the outset. However, the interpretation of a finding will be affected by its context. If a reported association is one of fifty which were examined, this should be clearly stated. We would probably view such an association (even with a small

P value) as generating a hypothesis which might be tested in future studies rather than evidence for or against the hypothesis.

2. Make analyses reproducible

During most practical sessions you have worked by typing a command, looking at the output, then correcting your command or typing the next command: you worked **interactively**.

It is sensible to start by working interactively, while getting a feel for the data and finding out what analyses are useful. But once you have worked out what you want to do, you should write programs consisting of these commands, so that it is easy to re-run your analysis. It is often helpful to save the list of commands you use when working interactively, and then edit them afterwards to retain only those that were useful. You can do this by typing, for example:

```
cmdlog using datacheck.do
```

before starting your analysis, and then

```
cmdlog close
```

when you have finished.

You can edit the series of commands you have generated either in a Word-processing package such as Word, or in the STATA do-file editor.

To rerun the commands saved in the file datacheck.do, and save the output in a file datacheck.log, type:

```
log using datacheck.log, text
do datacheck.do
log close
```

in the command line in STATA.

In STATA, a file containing a list of commands is known as a 'do' file. It is conventional to give these files the extension .do, e.g. datacheck.do to do data checking. It is often useful to write a series of 'do' files which carry out distinct parts of the analysis.

The importance of writing these 'do' files is that it ensures that you can **reproduce your analyses easily** if data are updated/corrected, or if you realise that you need to modify your analysis slightly, for example to control for an additional confounder. In practice your initial analysis will very often need to be redone, and in the long run saving the commands you use in a series of 'do' files will save a lot of time. As important, it provides a record of exactly what you did in your analysis.

3. Creation of a STATA data set

Most data will be entered into a database package such as Microsoft Access, or a spreadsheet package such as Excel, or a statistical package with facilities for data entry such as EpiInfo.

The best way to transfer data to a STATA format is to use a package such as **STAT-TRANSFER**. Many different file formats can be converted to a STATA format.

The dietsme2 data set has been provided as an Excel file, as well as a STATA file. We illustrate the use of STAT-TRANSFER to convert the dataset dietsme2.xls to a STATA dataset dietsme2.dta.

(i) To use STAT-TRANSFER, double-click on School Applications on the school network, and then double-click on Statistical Applications. Then double-click on STAT-TRANSFER 7 to use this package.

(ii) The top box is called 'Input File type'. You need to select Excel.

(iii) The next box is called 'File Specification'. To the right of the box is a button called 'Browse'. You need to click on Browse to select the file you want to convert. You need to select the directory h:\sme, and then select the file dietsme2.xls. Then click on 'Open'.

(iv) The next box is called 'Output file type'. You want to select STATA 8.

(v) The next box is called 'File specification'. You use this to choose where to save the transferred file, and the name of the transferred file. By default, STAT-TRANSFER will choose the same directory as the source file, and will give it the same name but with an extension .dta to show it is a STATA file. You can change the directory and file name if you wish, by editing the directory path and file name in the 'File Specification' box. But it is logical to use the default, which will be h:\sme\dietsme2.dta

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 6

Analysis of Unmatched Case-Control Studies

The objectives of this session are to be able to:

i) estimate odds ratios which control for the effect of a confounder in an unmatched case-control study
ii) test the null hypothesis OR=1, after controlling for a confounder
iii) decide whether a third variable modifies the effect of an exposure on a binary outcome
iv) test for a trend (eg a dose-response relationship) in the odds of a binary outcome, in a simple model without confounders.

For i–iii), Mantel-Haenszel techniques are used.

1. Introduction

In the last session we considered some of the theory underlying the interpretation of the odds ratio obtained from a case-control study, and performed some simple analyses of a single 2x2 table. Such analyses alone will usually be inadequate. We typically need to investigate the *joint* effects of two or more factors because of the possible presence of:

(i) Confounding, ie an apparent effect of exposure on disease being distorted by a confounding factor which is a risk factor for the disease *and* is associated with the exposure.

(ii) Interaction (effect modification), ie the effect of exposure on disease varying according to the level of another factor.

In addition, many exposures are not simple dichotomies. They may have several levels, or even be continuous. In such situations, categorising individuals as simply exposed or unexposed is wasteful of data.

More details on today's topics can be found on pages 136-156 of Breslow and Day, Volume 1 or in Chapter 5 (pages 85-136) of Kahn and Sempos. Chapters 18 and 20 of Clayton and Hills also examine these issues.

2. Adjusting the odds ratio for confounding factors (objective i)

The simplest technique for controlling confounding is to divide the sample into strata defined by levels of the confounder. If each stratum is homogeneous with respect to the confounding variable, the stratum-specific odds ratios (the odds ratio obtained from each table) cannot be biased by the effect of that confounding variable. In practice, there may be some residual confounding as subjects within strata will often not be identical with respect to the confounder.

Unless the estimates of the odds ratio obtained from each stratum differ markedly (see later), it is desirable to combine these estimates into a single, overall figure known as the *adjusted* or *corrected* odds ratio (or the odds ratio *controlled for...*). The motivation for this is twofold: first, the data within strata may be sparse rendering the individual estimates unreliable; second, we should always aim to summarise the data as succinctly as possible.

2.1 The Mantel-Haenszel estimate of the odds ratio (MHOR)

In general, the entries in stratum (table) j may be written as follows:

	Exposed	Unexposed	Total
Cases	D_{1j}	D_{0j}	D_j
Controls	H_{1j}	H_{0j}	H_j
	N_{1j}	N_{0j}	N_j

$j = 1,...,J$ where J represents the number of strata

The odds ratio for this table is:

$$OR = \frac{D_{1j}}{H_{1j}} \Big/ \frac{D_{0j}}{H_{01j}} = \frac{D_{1j} \times H_{0j}}{D_{0j} \times H_{1j}}$$

The obvious approach to forming a summary measure is to average the individual OR_j, but in preference to the simple average we calculate a weighted average which gives more weight to the strata with more data. A number of weighting schemes have been proposed, but the most widely used is that proposed by Mantel and Haenszel (1959):

$$w_i = \frac{D_{0j} \times H_{1j}}{N_j}$$

$$MHOR = \frac{\sum (w_j \times OR_j)}{\sum w_j} = \frac{\sum D_{1j} H_{0j} / N_j}{\sum D_{0j} H_{1j} / N_j} = \frac{Q}{R}$$

Note the similarity of this summary estimate to that used for comparing two sets of rates in a cohort study (appendix to session 2). See Appendix 3 in this session for an intuitive justification of these weights.

Example

In the Tanzanian study of HIV infection, a crude analysis of the association between schooling and HIV infection produced the following table:

	Educational Level		
	some formal education	none/ adult only	total
Cases	140 (74%)	49	189
Controls	311 (54%)	263	574
			763

Exercise 1 Work out the odds ratio for education as a risk factor for HIV, and write in the answer below.

OR =

χ^2 = 23.2, $p < 0.0001$

This result suggests that having been to school is associated with more than a doubling in the odds of HIV infection. This result is perhaps surprising and certainly worrying for those who believe in education.

A possible explanation for the observed asociation is confounding, and one possible confounding factor is age. One might hypothesize that women who have been more sexually active in the last 10 years — mainly in their 20s now — are those who are most likely to be HIV positive, and that younger people are more likely to have been to school.

Classifying women into four age groups (15-19,20-29,30-44,45-54), we can do a stratified analysis to test this.

Appendix 1 recaps the criteria which a confounder has to satisfy and what preliminary analyses can help to assess whether a factor is likely to be a confounder. First, we can see that age is associated with being a case, people in their twenties being over-represented among cases.

The percentage distribution of cases and controls by age (in years):

	15-19	20-29	30-44	45-54	total
cases	6.9	50.7	33.3	9.0	100 (n=189)
controls	16.7	33.4	33.4	16.4	100 (n=574)

The following table shows separate strata, and also shows that the proportion of controls who have schooling decreases from 81% in the youngest age group to 13% in the oldest age group. Thus there seems to be scope for confounding, as age is not a result of schooling or HIV status.

Age-stratified associations between schooling and HIV status

		Schooling				odds ratio	M-H weight
Age (years)		Yes		No	total		
15-19	Case	9	(69%)	4	13	0.52	2.8624
	Control	78	(81%)	18	96		
	total	87		22	109		
20-29	Case	82	(85%)	14	96	1.95	
	Control	144	(75%	48	192		
	total	226		62	288		
30-44	Case	44	(70%)	19	63	3.46	
	Control	77	(40%)	115	192		
	total	121		134	255		
45-54	Case	5	(29%)	12	17	2.85	
	Control	12	(13%)	82	94		
	total	17		94	111		

Exercise 2 Work out the Mantel-Haenszel weight for each stratum.

Using the Mantel-Haenszel formula we obtain a combined odds ratio of 2.29 which is slightly lower than 2.42 (the answer to Exercise 1). This calculation is shown in Appendix 2: the smaller strata have less weight. Thus age only seems to account for a small part of the negative association between schooling and HIV infection.

Normally we would do Mantel-Haenszel analysis by computer, but doing this example 'by hand' shows the principle on which it is based: weighting the stratum ORs according to the information each.

NOTE. This judgement about confounding is based on a comparison of the crude and adjusted estimates (2.42 and 2.29 respectively) *without* any statistical tests.

2.2 Testing the null hypothesis that the adjusted odds ratio = 1 (objective ii)

We may test the hypothesis that the true adjusted odds ratio equals 1.0 by considering the difference between the observed and expected number of exposed cases accumulated over all strata. An approximate test is provided by an extension of the Mantel-Haenszel chi-squared test for a single 2x2 table (previous session).

$$\chi^2_{MH} = \frac{U^2}{V} \qquad \text{on 1 df}$$

where $U = E D_{1j} - E E_{1j}$

and $V = \sum \frac{D_j \times H_j \times N_{0j} \times N_{1j}}{N_j^2 (N_j - 1)}$ $(= E\ V_j)$

When there is only one stratum this reduces to the formula presented in the last session for a single 2x2 table. Note the similarity of this formula to that for the stratified analysis of rates.

For the Mwanza example, the calculations for the test are as follows:

j	D_{1j}	E_{1j}	V_j
1	9	$10.38 = \frac{87 \times 13}{109}$	1.86
2	82	75.33	10.85
3	44	29.89	11.88
4	5	2.60	1.88
Total	140	118.21	26.47

$$\chi^2_{MH} = \frac{(140 - 118.21)^2}{26.47} = 17.94 \text{ on 1 df } (P<0.001)$$

After controlling for age, there is still strong evidence of an association between schooling and HIV infection, although the chi-squared statistic has been somewhat reduced.

2.3 Forming a confidence interval for the odds ratio

Several methods exist for forming an approximate confidence interval for the adjusted odds ratio. The most intuitive, and easiest to calculate, is the test-based confidence interval (Kahn & Sempos p117-120).

2.3.1 Test-based confidence interval. The idea is that the χ^2 statistic tells us how big the odds ratio is, compared to its standard error. More specifically, we treat it as the square of the ratio of the log-OR to its standard error:

$$\chi^2_{MH} = \left(\frac{\log(OR)}{\text{se}(\log(OR))} \right)^2$$

$$\text{se}(\log(OR)) = \frac{\log(OR)}{\sqrt{\chi^2_{MH}}}$$

The error factor (the number by which we divide and multiply the OR to get the 95% confidence interval) is then

$$\exp(1.96 \times \text{se}(\log(OR)))$$

Exercise 3 Work out the error factor for the MHOR from the stratified Mwanza data, and calculate the 95% confidence interval.

2.3.2 Confidence interval based on score variance This method is described by Clayton and Hills (p146 & 178). A 95% confidence interval for the odds ratio is found by dividing and multiplying MHOR by the following error factor:

$$EF = \exp(1.96 \times S) \qquad \text{where } S^2 = \frac{V}{QR}$$

Q and R are as defined as for the Mantel-Haenszel estimate of the odds ratio, and V is calculated as:

$$V = \Sigma V_j, \qquad V_j = \frac{D_j \times H_j \times N_{0j} \times N_{1j}}{N_j^2 (N_j - 1)}$$

Notice that V_j here is equivalent to the denominator (V) used in the score (χ^2) test in the previous session.

This method is the one used by the **mhodds** command in STATA. It gives an error factor of 1.483, and a 95% confidence interval of 1.54-3.40. For the test-based method (the answer to exercise 3), the results are similar: the error factor is 1.506 and the 95% CI is 1.52-3.45.

2.3.3 Which confidence interval to use? It is generally accepted that Cornfield's method, as presented by Gart (1982) is the most accurate (and is used by the **epitab** group of commands in STATA, except **tabodds** and **mhodds**), but it is an iterative method and difficult to calculate by hand. Moreover, the differences between the methods are generally unimportant (Kahn & Sempos p120-1). The test-based method is slightly less accurate when there is a large departure from the null hypothesis, but in these circumstances the loss of accuracy is less likely to be important. Kahn & Sempos conclude that the test-based method 'has been found useful over a fairly wide range of departures from the null hypothesis, and you need not hesitate to use it' (p120). In practice, the calculations will be done by computer and any of the above-mentioned methods is likely to be adequate except possibly in extreme circumstances (eg high odds ratios estimated from sparse tables), in which case Cornfield's method is recommended.

2.4 Interaction (objective iii)

We have not yet examined the data for the presence of *odds ratio interaction* (also known as *effect modification* or *heterogeneity of effect*). There is said to be interaction (effect modification) between age and schooling if the true odds ratios for HIV infection differ

between the age groups. Were this the case, it would be illogical to calculate a summary odds ratio; we should proceed no further than the calculation of stratum-specific odds ratios.

We have considered age as a confounder but it could be an effect modifier. It is possible that (for example) the nature of education has changed over the years such that, say, whereas education gives people opportunities for risky behaviour, the education in recent years may have sufficiently stressed the dangers of HIV to have deterred educated people from putting themselves at risk. If this is so we will find that the association between education and HIV differs for different age-groups, i.e that there is an interaction between age and education. The stratum-specific odds ratios in the table above seem to suggest different effects at different ages.

There are a number of different tests for heterogeneity of the odds ratios. STATA uses one which is based on the score statistic, and is run using the **mhodds** command with the option **by()**.

This yields

$$\chi^2 (3df) = 8.03, p=0.05$$

Thus there is some evidence of interaction, suggesting that schooling does not have such an adverse consequence among younger people. (We need to look again at the stratum-specific odds ratios to see which way round the interaction is tending.) We then have to decide whether the heterogeneity is great enough to invalidate the summary odds ratio.

As one reason for performing the test for interaction is to justify the calculation of a summary odds ratio, it may seem more sensible to apply it before, rather than after, deriving the Mantel-Haenszel odds ratio. However, that there is invariably some degree of interaction (odds are never *exactly* multiplicative), and the statistical test often has low power, so a non-significant result should not necessarily be interpreted as establishing the absence of interaction. On the other hand, if a test *does* have sufficient power to establish an interaction of low magnitude, some epidemiologists would ignore it to simplify the presentation of the data. Thus, it is worth examining the actual pattern of odds ratios: how different do they look, and is there any trend across strata? In general, when reporting an analysis, you should state the strategy used for detecting and dealing with interactions.

Note Statistical tests for interaction are based on assumptions of how the effects of different risk factors combine. For example, in the above models, when we test for an interaction of odds ratios, the null hypothesis (of no interaction) is that the odds ratios combine multiplicatively. If this null hypothesis were true, ie if the odds ration do combine multiplicatively, then a different null hypothesis of no *additive* interaction would *not* be true. Which null hypothesis is more reasonable to consider depends on the biological process. In general, statistical tests of interaction do not necessarily tell you about the presence or absence of biological interactions. For biological interaction to exist there has to be a causal mechanism whereby the outcome, given a particular level of one risk factor, depends on the level of a second risk factor. That mechanism may be additive rather than multiplicative. Rothman & Greenland Chapter 18 gives a detailed advanced level discussion of concepts of interaction.

Statistical interactions can easily be tested for within logistic regression models. In these models, a test of the null hypothesis that the effect (odds ratio) is the same in all strata is equivalent to testing an interaction term between the exposure (schooling) and the stratifying variable (age). The mechanics of testing for interaction/effect modification within logistic regression will be dealt with in later sessions.

3. A caution if the data are sparse

Some of the methods described are invalid if the data are sparse.

(i) the Mantel-Haenszel point estimate of the odds ratio is *not* invalidated by small sample sizes

(ii) the method we have given for forming a confidence interval for the odds ratio will be inaccurate if the overall sample size is small.

(iii) The Mantel-Haenszel test can legitimately be applied, save in extreme cases. See Kirkwood (page 102) for validity criteria.

4. Controlling for more than one confounder

It is possible to apply the above methods to control simultaneously for two or more confounders. For example, we can control for the woman's marital status (eg grouped into three categories: never married, married, and divorced/widowed) along with age, by forming the 3x4=12 strata corresponding to all combinations of marital status and age group. The drawback to this approach is that with many strata, the data may be spread so thinly that the approximate methods cease to be valid or data are wasted.

The alternative approach is to use a regression model. This will be considered in later sessions. In the analysis of a data set, we usually begin with the classical methods in order to get a feel for the data, considering at most two confounders at a time. This enables us to identify a subset of variables which we wish to consider further. After this, we apply the regression techniques which, although more powerful, are not as intuitive.

5. Exposure classified at three or more levels

So far in our analyses of the association between schooling and HIV infection, we have arbitrarily classified people as unexposed (never having been to school) or exposed (went to school). If, for the sake of argument, we had observed that women who had been to school had a lower risk of HIV infection (which we didn't), and we believed the link to be causal, we might have hypothesized that the longer the women went to school, the lower her risk of HIV infection. In fact, we observed that women who have been to school are at higher risk of HIV infection and that this increased risk is not explained by confounding with age. If schooling in some way affects behaviour or position in society in such a way as to *increase* risk of HIV infection, then we might expect to see risk of HIV infection increasing with

increased years of schooling. By categorising individuals as simply exposed or unexposed, we lose the opportunity of looking to see whether there is a "dose-response" effect.

Example (continued)

In the Mwanza study, data were collected on the number of years of schooling. To investigate whether duration of schooling is related to risk of HIV infection, we first define a *baseline* group and calculate an odds ratio for each exposure group relative to the baseline group. The 'unexposed' group is usually selected as the baseline group although sometimes the largest group is chosen as this gives the most stable odds ratio estimates.

	Years of schooling				
	None	1-3	4-6	7+	Total
Cases	49	24	110	6	189
Controls	263	51	255	5	574
	312	75	365	11	763
Estimated odds ratios	1.0 (baseline)	2.53 (OR_1)	2.32 (OR_2)	6.44 (OR_3)	

Hypothesis tests and interval estimates may be obtained for the odds ratio at each level of exposure compared with the baseline by applying the methods already described.

Two overall significance tests may be applied to data in this form. These tests examine the null hypothesis that the true odds ratio for all exposure levels is 1 ($OR_i = 1$ for all i) against two alternatives:

> **general association:** the true odds ratios for the different exposure levels are not all equal to 1 (for at least one i, OR_i is not equal to 1)
>
> **trend:** the true odds ratios increase (or decrease) with increasing exposure ($OR_i < OR_j$ for $i < j$).

The first of these alternatives is very general and tells us nothing about the nature of any association which we might observe. The second alternative is more specific and corresponds to there being a dose-response effect.

5.1 Testing for a general association

This test is performed using the standard χ^2 statistic for large contingency tables (see eg Statistics with Computing),

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where addition is over all cells in the table.

O_{ij} = observed frequency in cell in row *i* and column *j*, and

E_{ij} = expected frequency in cell in row i and column j

given by $\dfrac{\text{total for row } i \text{ x total for column } j}{N}$

This statistic is referred to the chi-squared distribution with (k-1) degrees of freedom, where k represents the number of exposure levels.

For the 2x4 table in our example this statistic is equal to 26.7. This value must be referred to the chi-squared distribution with 3 (4-1) degrees of freedom, giving a p-value of less than 0.0001. This indicates that our data provide strong evidence against the null hypothesis of no association and in favour of the alternative that risk of HIV infection is related to years of schooling - in some, undefined way.

5.2 **Testing for a trend in the odds** (objective iv)

To perform a test for trend for an increasing (or decreasing) odds of HIV with years of education, we assign a *score* to each exposure level. It is common to assign the scores 1, 2, 3, 4,....However, in an example like this, where the categories are based on actual numbers, it makes sense to use these numbers to define the scores eg the approximate midpoints of the categories: 0, 2, 5 & 9. In practice, though, the choice of score does not usually make much difference. We can compare the mean score between cases and controls by:

$$\chi^2_{trend} = \frac{(\bar{x}_1 - \bar{x}_0)^2}{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)} \text{ on 1 df}$$

where $\bar{x}_1$ = mean score for the cases
$\bar{x}_0$ = mean score for the controls
n_1 = total number of cases
n_0 = total number of controls
s = standard deviation of the score when cases and controls are combined

This is a z test comparing the average score between cases and controls. You may find it strange to use a z test on discrete rather than continuous data. Although not obvious from this formulation, in fact it is equivalent to the χ^2 test for trend described by Schlesselman (p200-203, following Mantel 1963), and is done by the **tabodds** command in STATA, and Epi Info's STATCALC module. The alternative hypothesis is that there is a linear trend between the scores and the log-odds of disease, ie that the odds are multiplied by a constant factor for each unit change in score. Hence the test is based on a parametric model and careless choice of scores can cause misinterpretation (Leuraud & Benichou 2001).

The test for trend in proportions (eg **ptrend** in STATA, Armitage & Berry 3rd ed p404, Altman p261-265, Fleiss p143-149) is slightly different to the test for trend in odds, but in practice will give similar answers.

The χ^2 for trend can be subtracted from the overall χ^2 (from the previous section) to give a χ^2 for departure from trend, on k-2 degrees of freedom. If this is small (eg the p value is more than 5 or 10%), it shows that the trend fitted (eg linear in log-odds, as shown here) explains satisfactorily all of the variation in the odds. If the χ^2 for departure from trend is large, there is evidence of other risk factors, or of a different kind of trend.

Example (continued)

	Years of schooling				
	None	1-3	4-6	7+	Total
Cases	49	24	110	6	189
Controls	263	51	255	5	574
	312	75	365	11	763
Score	1	2	3	4	

$\bar{x}_1 = 2.3862$ $\bar{x}_0 = 2.0035$, $s = 0.9677$

$$\chi^2 \text{ (trend)} = \frac{(2.3862-2.0035)^2}{0.96775 \times (1/189+1/574)} \quad \text{on 1 df}$$

$$= 22.24, \; p < 0.0001$$

This result indicates that our data provide strong evidence against the null hypothesis of no association, and in favour of the alternative that risk of HIV infection is related to years of schooling, with the risk increasing with years of schooling. Moreover, the χ^2 for departure from trend is 26.7-22.2=4.5 on 2 degrees of freedom, p>0.1. So a linear trend between education score and log-odds explains the variation in risk of HIV.

Of all the significance tests we have performed, the test for trend is the most powerful if such a dose-response relationship exists. This is because the test for trend makes the most efficient use of the data; the 2x2 analysis does not use the information on (eg) duration of exposure, and the test for a general association takes no account of the trend in the odds ratios.

5.3 Extensions to stratified data

The above approach to dealing with 2xk tables can be extended to stratified data. In an unstratified analysis, however, the odds ratios are consistent in the sense that $OR_{ij} = OR_{ik} \times OR_{kj}$ where OR_{ij} is the odds ratio for exposure level i against exposure level j, and i,j,k represent any three exposure levels. Unfortunately, this desirable property does not hold if the odds ratios have been adjusted for a stratifying variable.

We applied two tests, one against a very general alternative hypothesis, the other against the alternative of a trend in the odds ratio. There are analagous tests for stratified data. For a single 2xk table, the test for trend can be considered as a z-test comparing the mean 'score' between cases and controls. Similarly, the stratified test for trend (Breslow and Day, p149;

Schlesselman p203-6) can be thought of as a 2-way analysis of variance, assessing the effect of case-control status after allowing for the confounder. These tests are rarely used as regression methods (see later sessions) offer an alternative means for analysing such data.

Note on the association between education and HIV. A systematic review by Hargreaves and Glynn (2002) found an association, similar to that in the Mwanza study, in many studies done in Africa. The authors suggest that this association could result from the association between education and higher socio-economic status, and between the latter and the use of commercial sex workers by men. In women, the association could be explained by marriage partners tending to have similar socio-economic status. The authors find some evidence that the association between education and HIV has weakened in more recent studies, perhaps due to an increased component of HIV-related health education in schools.

6 Summary

In an unmatched case-control study, we can adjust an odds ratio for a confounder by a weighted average of the stratum-specific odds ratios. This is called the Mantel-Haenszel odds ratio (MHOR).

We can test the null hypothesis that the adjusted odds ratio (MHOR) equals 1 by a χ^2 test.

A confidence interval for the MHOR can be constructed by a test-based or other method.

We can use another χ^2 test to help decide whether a third variable modifies the effect of an exposure on the outcome (ie whether there is an interaction).

A trend between the odds of the outcome and the level of exposure (eg a dose-response relationship) can be tested for by another χ^2 test.

References

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, Chapman and Hall.

Armitage, P. and G. Berry (1994). *Statistical Methods in Medical Research*, third edition. Oxford, Blackwell Scientific Publications.

Breslow, N. E. and N. E. Day (1980). *Statistical Methods in Cancer Research: Volume 1 - The Analysis of Case-control Studies*. Lyon, International Agency for Research on Cancer.

Clayton, D. and M. Hills (1993). *Statistical Models in Epidemiology*. Oxford, Oxford University Press.

Dayal, H. H. (1978). On the desirability of the Mantel-Haenszel summary measure in case-control studies of multifactor etiology of disease. *American Journal of Epidemiology* 108(6): 506-511.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York, John Wiley and Sons.

Gart, J. J. and D. G. Thomas (1982). The performance of three approximate confidence limit methods for the odds ratio. *American Journal of Epidemiology* 115(3): 453-470.

Hargreaves, J. R. and J. R. Glynn. (2002). Educational attainment and HIV-1 infection in developing countries: a systematic review, *Tropical Medicine and International Health,* 7(6): 489-498.

Kahn, H. A. and C. T. Sempos (1989). *Statistical Methods in Epidemiology*. New York and Oxford, Oxford University Press.

Kirkwood BR. Essentials of Medical Statistics. Oxford: Blackwell Scientific Publications, 1988.

Leuraud K, Benichou J. A comparison of several methods to test for the existence of a monotonic dose-response relationship in clinical and epidemiological studies. *Statistics in Medicine* 2001;**20:**3335-3351.

Mantel, N. and W. H. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22: 719-748. (reprinted in Buck, C., A. Llopis, E. Nájera and M. Terris, Eds. (1988). *The Challenge of Epidemiology: Issues and Selected Readings*, pages 533-553. Scientific Publication Number 505. Washington DC, Pan American Health Organization.)

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* 58: 690-700.

Rothman, K. J. and S. Greenland (1998). *Modern Epidemiology*, second edition. Philadelphia, Lippincott-Raven Publishers.

Schlesselman, J. J. (1982). *Case-control Studies: Design, Conduct, Analysis*. New York, Oxford University Press.

Stata website: http://www.stata.com/support/faqs/stat/trend.html

Appendix 1: Deciding whether a factor may be a confounder

In some situations prior experience tells you that a factor is a risk factor for the outcome and is associated with your exposure.

However, where you are less familiar with the subject, you can undertake some preliminary analyses to see whether a factor is likely to be confounding the true association between the exposure and outcome.

1) Confounding factor as a risk factor for the disease (or a good proxy for a cause)
An observed association in your study between a candidate confounding factor and the outcome is neither a necessary nor a sufficient condition for the candidate factor to be a true risk factor for the outcome. Rothman and Greenland give some counter-examples (pp 121-122). Nevertheless, in the absence of other information you can tabulate the candidate factor by outcome and see whether the prevalence of exposure to this factor differs between cases and controls. (See the example for age on p6.3.)

2) Confounding factor associated with the exposure
The association must be present in the source population from which cases and controls are drawn. If the control series is large this can be assessed by cross-tabulating the factor by exposure among controls. For example:

Age Group		Schooling			
		Yes		No	Total
15-19	Control	78	(81%)	18	96
20-29	Control	144	(75%)	48	192
30-44	Control	77	(40%)	115	192
45-54	Control	12	(13%)	82	94

Pearson χ^2 on 3 df = 142.1284 P<0.001

3) A confounding factor must not be affected by the exposure or outcome.

This means it must not be on the causal pathway between exposure and outcome, e.g. plasma cholesterol levels are on the causal pathway between a diet high in saturated fats and cardiovascular disease. Also it must not be a consequence of the outcome. For example, abstinence from alcohol which results from feeling ill could not be a confounder for the relation between exercise and angina.

Even if a factor fulfils these three criteria it may turn out not to confound the association between your main exposure and outcome.

If is does confound, it may make the odds ratio appear further from 1.0 than it should be (inflate) or closer to 1.0 (deflate).
A confounder inflates the odds ratio if either

- The confounder is positively associated with both exposure and outcome and the fundamental association between exposure and outcome is positive (an adverse exposure for a disease). For example smoking as a confounder for measuring effect of coffee on cancer of the pancreas (smokers more likely to drink coffee and more likely to have this cancer).

Or

- The confounder is positively associated with one of exposure/outcome and negatively associated with the other and the fundamental association is negative (a protective exposure). For example education as a confounder for measuring effect of vaccination on TB (educated more likely to have vaccination and less likely to have TB).

Conversely, the crude odds ratio will deflate if either

- The confounder is positively associated with one of exposure/outcome and negatively associated with the other and the fundamental association between exposure and outcome is positive (an adverse exposure for a disease). For example, gender as a confounder for measuring effect of being sedentary on having a myocardial infarction (men less likely to be sedentary and more likely to have MI).

Or

- The confounder is negatively associated with both exposure and outcome and the fundamental association is negative (a protective exposure). For example urban/rural as a confounder for measuring effect of vaccination on TB (rural people less likely to get vaccinated and less likely to get TB).

Appendix 2 Calculations for Mantel-Haenszel estimates adjusted for a confounding variable: can complete in own time as a revision exercise.

Associations between education and HIV infection adjusted for age

1. Calculation of Mantel-Haenszel weights for each stratum and hence MHOR

Stratum j	Weight w_j	OR_j	$w_j \times OR_j$
1 (15-19yrs)	78x4/109 = 2.8624	0.5192	1.4862
2 (20-29yrs)	144x14/288 = 7	1.9524	13.6666
3 (30-44yrs)	77x19/255 = 5.7373	3.4586	19.8430
4 (45-54yrs)	12x12/111= 1.2973	2.8472	3.6937
	Σ=16.8970		Σ=38.6898

MHOR = 38.6898/16.8970 = 2.29

2. Alternative formulation

$Q = 3(D_1j\ H_0j/Nj)$ = (9x18)/109 + (82x48)/288 + (44x115)/255 + (5x82)/111

$R = 3(D_0j\ H_1j/Nj)$ = Σw_j = (4x78)/109 + (14x144)/288 + (19x77)/255 + (12x12)/111

And MHOR = Q/R

This should yield the same result

3. Calculation of the confidence interval for the MHOR (see section 2.2)

Stratum j	$V_j = \frac{D_j \times H_j \times N_{0j} \times N_{1j}}{N_j^2 (N_j - 1)}$
1 (15-19yrs)	$\frac{13 \times 96 \times 87 \times 22}{109 \times 109 \times 108} = 1.86157$
2 (20-29yrs)	10.8494
3 (30-44yrs)	
4 (45-54yrs)	
	V =

Error factor, EF = exp(1.96 S)
Where $S^2 = V/QR$

Lower end of the 95% confidence interval = MHOR/EF
Upper end of the 95% confidence interval = MHOR x EF

Appendix 3 Justification of the Mantel-Haenszel weights

Q: Why use the Mantel-Haenszel weights instead of other possibilities. In particular, why do only two cells seem to contribute: the unexposed cases and the exposed controls, not the other two?

A If we consider the *lack* of the risk factor as the exposure (rather than its presence), then, if we rearrange the table, we see that the other two cells are included rather than the original two. If the 0 and 1 subscripts are swapped in the formula for the MHOR, then we get 1/ the original odds ratio, as we should.

Q: OK, but still, why that particular form for the weights?

A: In fact the rationale for the weights is not totally exact in the original paper (Mantel & Haenszel 1959, Dayal 1978), where they are proposed as a 'compromise' between weighting by precision and by 'importance'. 'Importance' of an odds ratio is said to be greater if it applies to a larger population, or if it multiplies a larger baseline odds, although it is not quantified precisely. However, we can make an heuristic argument by rearranging the formula as follows.

$$
\begin{aligned}
w_j &= \frac{D_{0j} H_{1j}}{N_j} \\
&= \frac{D_{0j} H_{1j}}{D_j + H_j} \\
&= \frac{D_j H_j}{D_j + H_j} \frac{D_{0j}}{D_j} \frac{H_{1j}}{H_j} \\
&= \left(\frac{1}{H_j} + \frac{1}{D_j} \right)^{-1} \frac{D_{0j}}{D_j} \frac{H_{1j}}{H_j}
\end{aligned}
$$

Now the weight is split into three factors.

For the first part (raised to the power -1) to be large, both the number of controls (H_j) and the number of cases (D_j) must be large. Moreover, for a fixed stratum total, the weight is largest for equal numbers of cases and controls. This corresponds to the fact that greatest power is obtained for a given sample size by having equal numbers in the two groups being compared. Having a big overall stratum size is OK but the effect on the weight is diluted if the balance of cases and controls is uneven.

The second and third factors are the proportion of cases which are unexposed, and the proportion of controls which are exposed. Here we can reason as follows. There is no point having a stratum in which everyone is exposed, or everyone is unexposed. Having the weight proportional to the product of these two means that a large weight is achieved if, overall, there is substantial proportion of both exposed and unexposed. (And, remember, as explained above, that, if we prefer, we can put in the cells the other way around (ie exposed cases and unexposed controls); this does not affect the first factor, and overall will give us 1/ the original odds ratio.)

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 6 PRACTICAL

Analysis of unmatched case control studies

By the end of this practical students should be able to:

(i) use the STATA commands **tabulate** and **mhodds** to obtain an odds ratio and its confidence interval for 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.

(ii) use the **mhodds** command to investigate whether the effect of the main exposure of interest depends on value of a second variable (interaction).

This session's practical uses the dataset from Mwanza, Tanzania on HIV infection among women. You will find this in the data file called MWANZA.DTA in the directory u:\download\teach\sme.

1. Use the STATA commands **gen** and **recode** to create two new variables ed2 and age2 with the following categories:

   ed2 1=none/adult only 2= 1 or more years
   age2 1= 15-19, 2= 20-29, 3= 30-44, 4 = 45+

   Check that you have created the variables correctly.

2. Use **tabulate** to reproduce the first 2x2 table of the lecture. How does the table on the screen differ from that in the lecture notes? Try the following commands:

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

   What is the difference between these two commands? How does their output compare with the odds ratio given in the lecture notes?

   To test the null hypothesis of no association use the command

```
tabulate case ed2, chi exact
```

   (The option **exact** gives Fisher's exact test: is it necessary to have this in addition to the χ^2?)

   What do you conclude?

3. Use **tabulate** prefixed with '**by age2:**' to obtain the tables of HIV infection by education (**ed2**) stratified for age (you will have to **sort** by **age2** first). Use STATA to find the odds ratio for HIV infection comparing those with and without education within

each stratum. This is done using `mhodds` with the option `, by(age2)`. If appropriate, produce a summary estimate of the odds ratio adjusted for age.

4. Investigate whether religion (`rel`) confounds the association between schooling (`ed2`) and HIV infection (`case`). Warning: `rel` has a code 9 for missing values, which we suggest you set to system-missing (.).

   To confine the adjusted MHOR estimate to those cases with known religion, type:

   ```
   mhodds case ed2 if rel!=., by(rel) c(2,1)
   ```

5. In the lecture we saw that there was evidence of a dose-response effect of years of schooling on HIV infection. Use `tabodds` to perform a test for trend. What can you learn from this analysis? Repeat the same analysis using

   ```
   mhodds case ed
   ```

   What does the estimate of the odds ratio from the `mhodds` command represent?

6. To show that treatment of missing values can matter.

   The number of sexual partners (`npa`) is unknown for 28 people (code 9). Suppose we set these to system missing:

   ```
   recode npa 9=.
   ```

   Use STATA to give a crude estimate of the association between schooling and HIV infection and one adjusted for number of sexual partners.

   ```
   mhodds case ed2
   mhodds case ed2, by(npa)
   ```

   These estimates include missing values of `npa`

   Now exclude missing values.

   ```
   mhodds case ed2 if npa!=.
   mhodds case ed2 if npa!=., by(npa)
   ```

   Compare the two sets of answers.

7. OPTIONAL

   We might expect an increasing risk of HIV infection with number of sexual partners. Carry out a test for trend using `npa`. (Decide how you want to treat missing values first). Estimate the odds ratio for each increase in category of number of partners.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 7

LIKELIHOOD

Objectives

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

(i) Explain qualitatively how likelihoods can be used to obtain parameter estimates;
(ii) Explain qualitatively how likelihood ratios can be used to obtain supported ranges/confidence intervals;
(iii) Explain qualitatively how likelihood ratios can be used to examine/test hypotheses about a parameter of interest.

1. Models

All statistical analyses are based probability models, usually involving one or more parameters. In elementary statistics the model is often not made explicit, but it is there. Some familiar situations are described below in terms of statistical models.

Follow-up of a cohort over a fixed interval of time. Each subject in the cohort has the same risk of disease π and the outcomes for the different subjects are statistically independent. The probability model is the binomial distribution with parameter π, the risk.

A prevalence study. Each subject in the sample has the same probability π of exhibiting the disease or trait under study and the outcomes for the different subjects are statistically independent The probability model is the binomial distribution with parameter π, the prevalence.

Follow-up of a cohort over varying intervals of time. Each subject in the cohort has the same rate of disease λ, which is constant over time. The outcomes for the different subjects are statistically independent. The probability model is the Poisson distribution with parameter λ, the rate.

2. Likelihood

One use of models is to allow us to make statements about the value of a parameter (or parameters) of interest based on past observations (*data*). This process is called *estimation* and the most important general approach to it is called *likelihood.* In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data. However, the analysis is only scientifically useful if we can generalize the findings, and such generalization must imply a model. Although the formal machinery of modelling and estimation may seem heavy handed for simple analyses, an understanding of it is essential to the development of methods for more difficult problems.

Likelihood is a measure of the *support* provided by a body of data for a particular value of the parameter of a probability model. It is calculated by working out how probable our observations would be if the parameter were to have the assumed value. The main idea is that parameter values

which make the data more probable are better supported than values which make the data less probable. In this session we develop this idea within the framework of the binomial model.

2. Likelihood in the binomial model

The outcomes observed in a small study in which 10 subjects are followed up for a fixed time period are:

F F S F S S S F S S

There are two possible outcomes for each subject: *failure*, such as the development of the disease of interest, or *survival*. We adopt the binomial probability model for the outcome for each subject in which failure has probability π and survival has probability $1 - \pi$. To calculate the probability of occurrence of this result we simply multiply the probabilities of the individual outcomes:

$$\pi \times \pi \times (1 - \pi) \times \ldots . \times (1 - \pi) = (\pi)^4 (1 - \pi)^6$$

This expression can be used to calculate the probability of the observed study result for any specified value of π. For example, when $\pi = 0.1$ this expression takes the value

$$(0.1)^4 \times (0.9)^6 = 5.31 \times 10^{-5}$$

and when $\pi = 0.5$ it takes the value

$$(0.5)^4 \times (0.5)^6 = 9.77 \times 10^{-4}.$$

The results of these calculations show that the probability of the observed data is greater for $\pi = 0.5$ than for $\pi = 0.1$. In statistics this is often expressed by saying that $\pi = 0.5$ is more likely than $\pi = 0.1$, meaning that a value of 0.5 is better supported by the data than 0.1. In everyday use the words probable and likely mean the same thing, but in statistics the word likely is used in this more specialized sense.

Exercise 1

Is $\pi = 0.4$ more likely than $\pi = 0.5$?

We obtained the expression

$$\pi^4 (1-\pi)^6$$

by considering the probability of obtaining the observed data, but when we use this expression to assess the amount of support for different values of π it is called a *likelihood*. More generally, if we observed D failures in N subjects, the likelihood for π would be

$$\pi^D (1-\pi)^{N-D}.$$

Returning to our numerical example, Fig. 1 shows how the likelihood varies as a function of π. The value $\pi = 0.4$ gives a likelihood of 11.9×10^{-4}, which is the largest which can be achieved. This

value of π is called the *most likely value* or, more formally, the *maximum likelihood estimate* of π. It coincides with the observed proportion of failures in the study, 4/10.

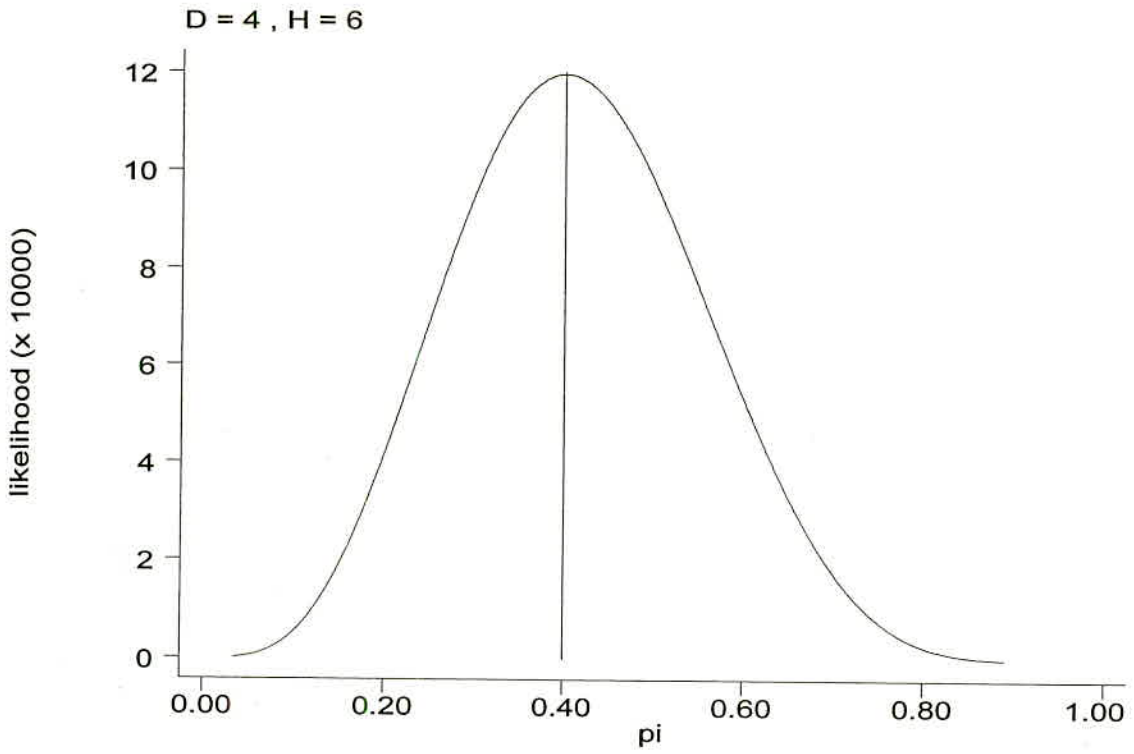


Figure 1. The likelihood for π based on a split of 4:6

3. The supported range for π

The most likely value for π is 0.4, with likelihood 11.9×10^{-4}. The likelihood for any other value of π will be less than this. How much less is measured by the *likelihood ratio*, which takes the value 1 when $\pi = 0.4$ and values less than 1 for any other values of π. This provides a more convenient measure of the degree of support than the likelihood itself. It can be used to classify values of π as either supported or not according to some critical value of the likelihood ratio. Values of π with likelihood ratios above the critical value are reported as "supported", and values with likelihood ratios below this critical value as "not supported". The *supported range* for π is the set of values of π with likelihood ratios above the critical value. The choice of the critical value is a matter of convention.

For our observation of 4 failures and 6 survivors, the likelihood ratio for different values of π is shown in Figure 2. We have used the number 0.1465 for the critical value of the likelihood ratio. The range of supported values for π is rather wide in this case: from 0.15 to 0.70. These values were obtained from the graph, as illustrated - we shall be describing more convenient approximate methods for their computation later. For any choice of critical value the width of the supported range reflects the uncertainty in our knowledge about π. The main thing which determines this is the quantity of data used in calculating the likelihood. For example, if we were to observe 20 failures in 50 subjects, the most likely value of π would still be 0.4, but the supported range would be narrower (see Figure 3).

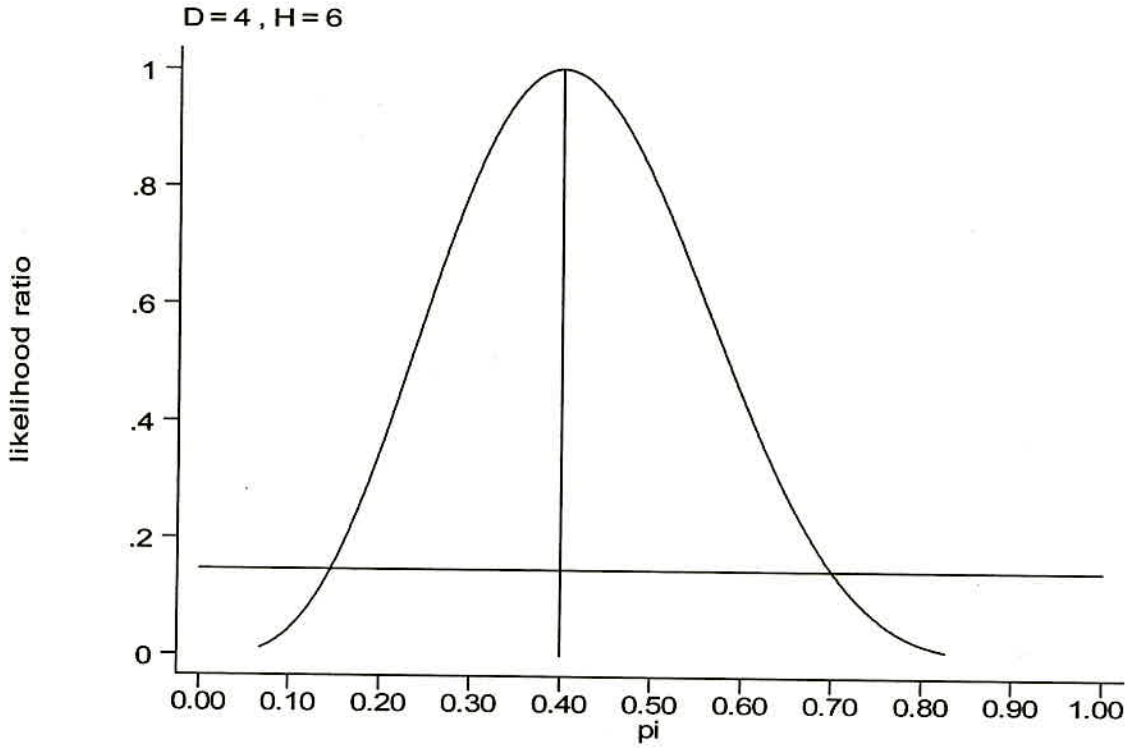


Figure 2. The likelihood ratio, with cut-off of 0.1465 indicated, for π based on a split of 4:6

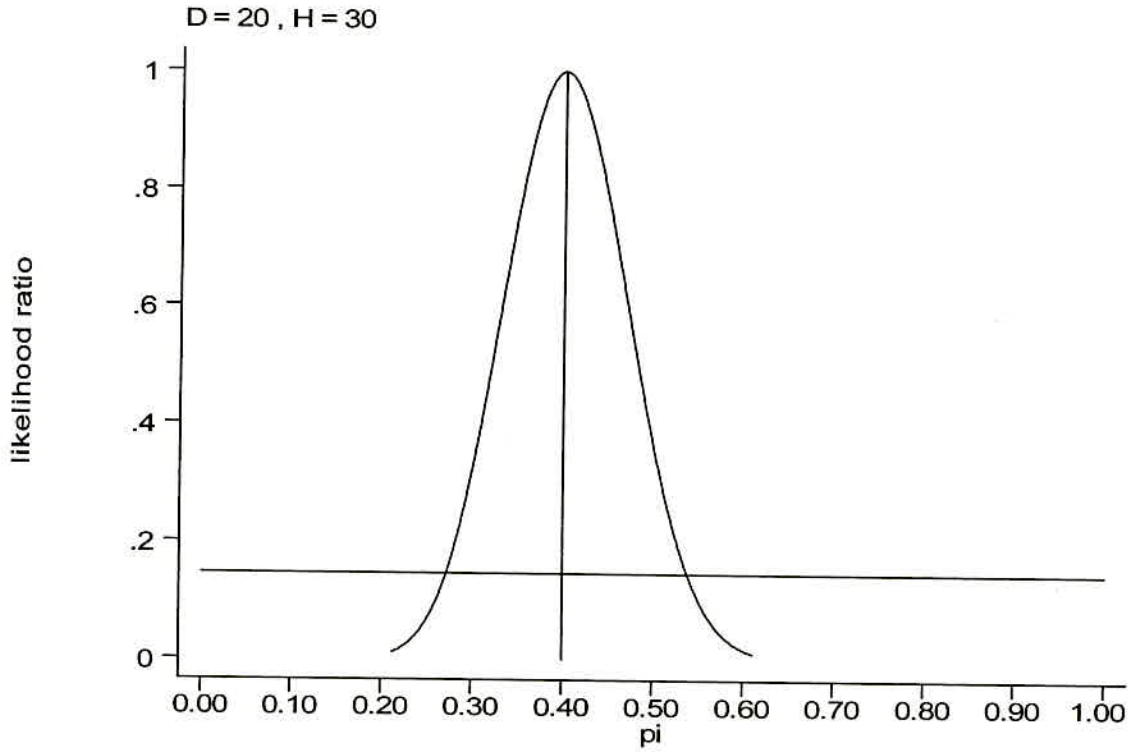


Figure 3. The likelihood ratio, with cut-off of 0.1465 indicated, for π based on a split of 20:30

4. Gaussian likelihood

When a quantitative outcome has a Gaussian (or normal) distribution with mean μ and standard deviation σ, the probability of observing a value close to x is proportional to

$$\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

This is also the likelihood ratio for μ based on a single observation x; it takes its maximum value when μ = x and this maximum value is 1.

If we had just one sample, x, from a normal distribution whose mean μ was unknown but whose standard deviation σis known, we would take x as our estimate of μ and we would calculate 95% confidence limits for our estimate of μ using the following formula:

$$\mu_{upper/lower} = \text{point estimate} \pm 1.96 \times \text{standard error}$$

In general the standard error is $\sigma/\sqrt{N}$. In our simple example N = 1 and so the formula becomes

$$\mu_{upper\,/\,lower} = x \pm 1.96\sigma$$

This can be rewritten as:

$$\left(\frac{x-\mu_{upper\,/\,lower}}{\sigma}\right) = \pm 1.96$$

which gives

$$\left(\frac{x-\mu_{upper\,/\,lower}}{\sigma}\right)^2 = 3.8414$$

which in turn gives

$$\left(-\frac{1}{2}\left(\frac{x-\mu_{upper\,/\,lower}}{\sigma}\right)^2\right) = -1.9207$$

and

$$\exp\left(-\frac{1}{2}\left(\frac{x-\mu_{upper\,/\,lower}}{\sigma}\right)^2\right) = 0.1465$$

So the upper and lower 95% confidence limits for μ are the values of μ which satisfy this equation. But the left hand side of this equation is just the likelihood ratio. I.e. the upper and lower 95% limits for μ are those values which produce a likelihood ratio of 0.1465.

In the general case where there are N observations of x with mean $\bar{x}$ and standard error $\sigma/\sqrt{N}$ the Gaussian likelihood for μ is

$$\exp\left(-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right)^2\right)$$

and the 95% confidence limits are given by

$$\mu = \bar{x} \pm 1.960\sigma/\sqrt{N}$$

Figure 4 shows the likelihood ratio for μ based on a sample of 20 observations of blood pressure for which the mean was 128 and the standard deviation was 13.9. The standard error (S) is 13.9/√20 =3.11, which gives 95% confidence limits

128-1.96 x 3.11 = 121.9 , 128+1.96 x 3.11 = 134.1.

The way we have chosen the critical value for the likelihood is to come down the vertical scale to the point (0.1465), which corresponds to 1.96 standard errors either side of 128. Using this critical value means that likelihood based 95% confidence limits will be exactly the same as conventional 95% confidence limits when the likelihood curve has an exactly Gaussian shape. Choosing a critical value 0.2585 would be equivalent to using 90% confidence intervals.

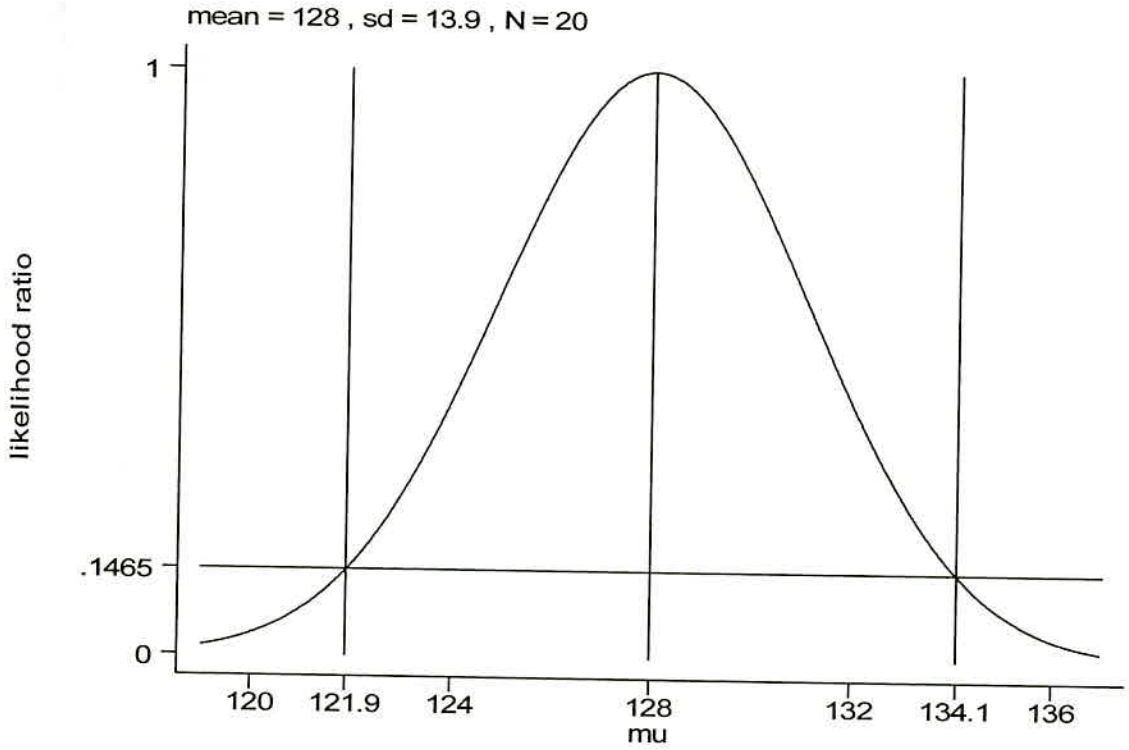


Figure 4. Critical value of the likelihood for a 95% confidence interval

6. Null values

In some situations there is one particular value of the parameter which is of particular interest (e.g. risk ratio = 1 which corresponds to no effect of the exposure being investigated). This value is called the null value, or *null hypothesis*. In the past considerable emphasis was placed on the need to disprove (or reject) the null hypothesis before claiming positive findings, and the procedures which are used to this end are called *hypothesis tests* or statistical significance tests. However, this emphasis on rejecting or not the null hypotheses has led to widespread misunderstanding and misreporting in the medical research literature. In epidemiology, which is not an experimental science, the usefulness of the idea has been particularly questioned. Undoubtedly the idea of statistical significance testing has been overused, at the expense of the more useful procedures for *estimation* of parameters, but it is still useful. A null hypotheses is a simplifying hypothesis and measuring the extent to which the data are in conflict with it remains a valuable part of scientific reasoning.

7. Hypotheses concerning a single parameter

In recent years there has been a trend away from a making a straight choice between rejecting or not the null hypothesis. Instead, the degree of support for the null hypothesis is measured, for example using the log likelihood ratio at the null value of the parameter. To illustrate the general idea of using the log likelihood ratio to measure the support for a null value, we shall consider a simple experiment in which 15 subjects are asked to choose between two treatments, A and B. Suppose that 12 choose A and 3 choose B. The model is that the probability of choosing A is π, the probability of choosing B is $1-\pi$, and the outcomes are independent. The null value of π is 0.5, corresponding to the two treatments being equally preferable. The likelihood for π is

$$\pi^{12}\,(1-\pi)^{3}$$

and the most likely value for π is 12/15=0.8. The value of the likelihood when π=0.8 is

$$(0.8)^{12}\,(0.2)^{3} = 5.4976 \times 10^{-4}\,.$$

The likelihood for π=0.5 is

$$(0.5)^{12}\,(0.5)^{3} = 3.0518 \times 10^{-5}\,,$$

so the likelihood ratio for π=0.5 is

$$\frac{3.0518 \times 10^{-5}}{5.4976 \times 10^{-4}} = 0.0555.$$

This likelihood ratio, along with the likelihood ratios for other values of π are shown in Figure 5; clearly π = 0.5 is well outside the 95% range, and is not well supported by the data, but as a measure of the degree of support the likelihood ratio lacks an intuitive appeal.

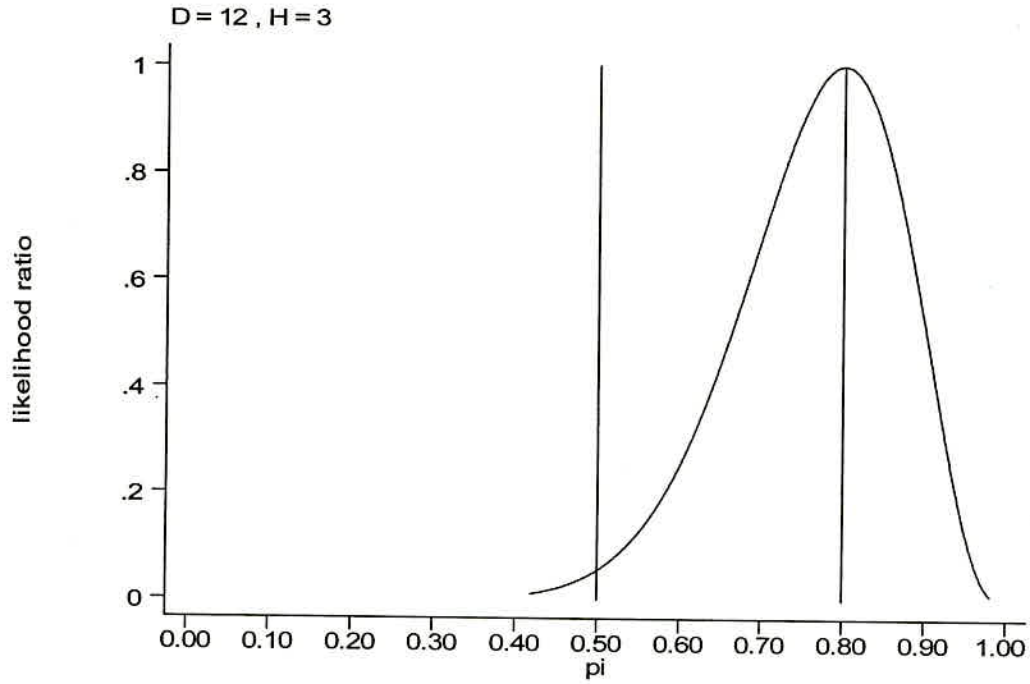


Figure 5. The likelihood ratio for π, based on the split 12:3

A commonly used measure is provided by the p-value obtained by working out the probability that a χ^2 distribution with 1 degree of freedom exceeds the value of:

$$-2 \times \log_e \text{likelihood ratio}$$

The rationale for this is that

$$\text{normal likelihood ratio} = \exp\left(-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right)^2\right)$$

$$\log_e(\text{likelihood ratio}) = \left(-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right)^2\right)$$

$$-2 \times \log_e(\text{likelihood ratio}) = \left(\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right)^2\right)$$

The right hand side of this equation is the square of a random variable with a N(0,1) distribution – the definition of the χ^2 distribution with 1 degree of freedom.

In this case: $-2 \times \log_e(0.0555) = 5.78$, $p=0.016$.

Like the choice of critical value of the likelihood ratio for the 95% confidence interval this equivalence between the log likelihood ratio and the p-value is exactly true for the Gaussian distribution, and will be approximately true for other distributions providing we have a reasonable amount of data.

Summary

- likelihood quantifies the extent to which data support different values of the parameter of interest

- to obtain point estimate of parameter, choose the parameter value which maximises the likelihood

- can also obtain a supported range (confidence interval) by choosing a cut-off for the likelihood ratio and perform statistical significance (hypothesis) tests using the likelihood ratio test.

Appendix Poisson likelihood

The likelihood for a rate λ based on D cases and Y person years is

$$\lambda^D \exp(-\lambda Y).$$

The most likely value of λ is D/Y. When D= 7 and Y=500, for example, the likelihood is

$$\lambda^7 \exp(-500\lambda),$$

and the most likely value of λ is 7/500= 0.014 or 14/1000 person-years. The likelihood ratio curve is shown in the Figure. The supported range is from 6.0 to 27.0 per 1000.

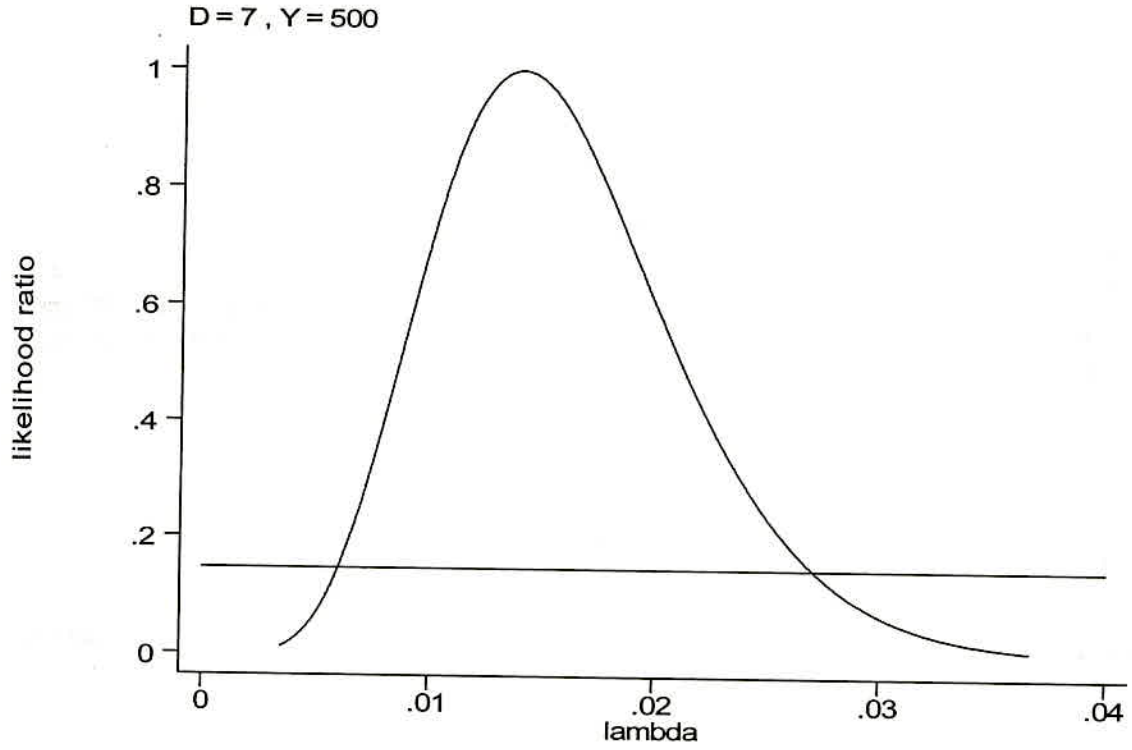


Figure. The likelihood ratio for λ based on 7 events in 500 person years

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 8

APPROXIMATE LIKELIHOODS

Objectives

By the end of the session the student will be able to:

(i) Explain why we usually prefer to work with the log likelihood ratio,
(ii) Explain in qualitative terms the use of the quadratic approximation to derive approximate confidence intervals,
(iii) Explain in qualitative terms the difference between likelihood ratio, Wald and score tests,
(iv) Explain why we usually work with the log(rate) and log(odds) rather than directly with the rate and risk.

1. The log likelihood

In the previous session we saw how likelihood theory can be used to obtain parameter estimates and how we can use the likelihood ratio to obtain supported ranges (confidence intervals) and to perform statistical tests. However, for reasons that will become clear, it is generally more convenient to use the natural logarithm (to base e) of the likelihood in place of the likelihood itself. The log likelihood for π, based on 4 failures and 6 survivors, is the logarithm of

$$\pi^4 (1-\pi)^6$$

which is

$$4 \log_e(\pi) + 6 \log_e(1-\pi).$$

The log likelihood takes its maximum at the same value of π as the likelihood, namely $\pi = 0.4$, so its maximum is

$$4 \log_e(0.4) + 6 \log_e(0.6) = -6.730.$$

Exercise 1

Calculate the log likelihood when $\pi = 0.5$, and verify that this is less than the log likelihood for $\pi = 0.4$.

The likelihood ratio for $\pi = 0.5$ is

$$\frac{(0.5)^4 (0.5)^6}{(0.4)^4 (0.6)^6} = 0.8176,$$

so the log likelihood ratio is

$$\log_e(0.8176) = -0.2014.$$

This can also be obtained by subtracting the log likelihood for $\pi = 0.4$, which is

$$4\log_e(0.4)+6\log_e(0.6) = -6.7301$$

from the log likelihood for $\pi = 0.5$, which is

$$4\log_e(0.5)+6\log_e(0.5) = -6.9315$$

to give

$$-6.9315 -(-6.7301) = -0.2014.$$

Graphs of both the likelihood ratio and the log likelihood ratio are shown in Figure 1. The supported range for π can be found from the left hand graph by finding those values of π for which the likelihood ratio is 0.1465, or from the right hand graph by finding those values of π for which the log likelihood ratio is equal to $\log_e(0.1465) = -1.921$. Both approaches give the same result.

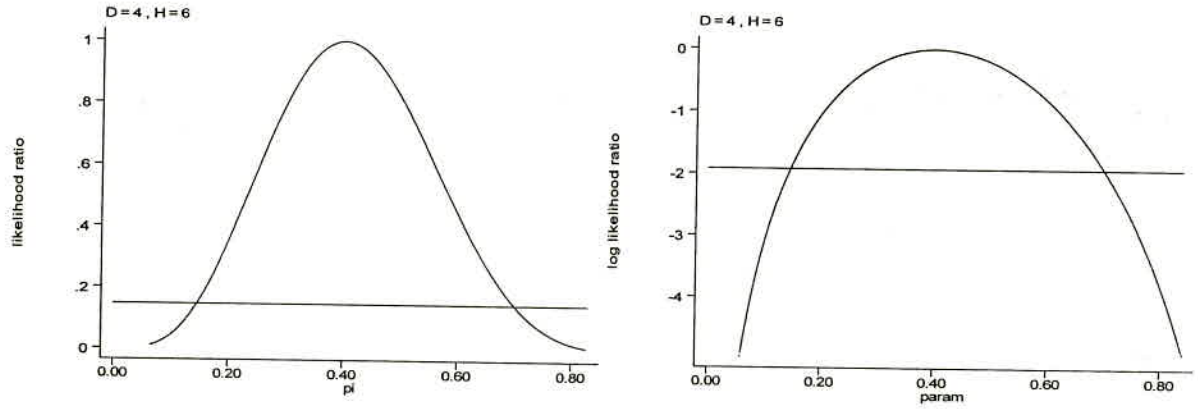


Figure 1 The likelihood ratio and log likelihood ratio for π

2. Formulae for likelihoods and log likelihoods

The risk parameter (π)

In general, the likelihood for π, when D subjects fail and N-D survive, is

$$\pi^D (1-\pi)^{N-D},$$

and the log likelihood is

$$D\log_e(\pi) + (N-D)\log_e(1-\pi).$$

Both expressions take their maximum value when $\pi = D/N$, the observed proportion of subjects who failed (see Figure 1 where the most likely value is $\pi = 0.4$).

The odds parameter (Ω)

The likelihood for Ω, the odds parameter, is obtained from the likelihood for π by replacing

$$\pi = \frac{\Omega}{1+\Omega}, \qquad 1-\pi = \frac{1}{1+\Omega}$$

which gives

$$\left(\frac{\Omega}{1+\Omega}\right)^{D}\left(\frac{1}{1+\Omega}\right)^{N-D} = \frac{\Omega^{D}}{(1+\Omega)^{N}}$$

as the likelihood for Ω. The corresponding log likelihood is

$$D\log_e(\Omega) - N\log_e(1+\Omega).$$

The most likely value of Ω is D/(N-D).

The mean of the Gaussian (normal) distribution (μ)

The Gaussian likelihood for μ based on a single observation x from a Gaussian distribution with unknown mean μ and known standard deviation σ is

$$\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}\right)$$

and the corresponding log likelihood is

$$\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}\right)$$

Both expressions take their maximum value when $\mu = x$. In terms of the unknown parameter μ, the log-likelihood is a quadratic expression; i.e. an expression containing μ and μ^2, but no other powers of μ.

3. Approximating the log likelihood

The values of the parameters at which the likelihood ratio is 0.1465, or the log likelihood ratio is -1.921, must be found, in general, by trial and error. However, the shapes of most likelihoods are *approximately* Gaussian, which means the log likelihoods will be approximately quadratic. Because we can easily solve quadratic equations, this fact can be used to derive simple formulae for approximate supported ranges/confidence intervals. Methods based on quadratic approximations to the log likelihood are particularly important because the quadratic approximation becomes closer to the true log likelihood as the amount of data increases.

For the Gaussian distribution, 95% confidence limits can be obtained by finding the values of μ which solve the equation:

$$\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) = -1.921$$

We have seen that these limits correspond are $\mu_{upper/lower} = x \pm 1.96\sigma$

Consider a general likelihood for the parameter, Θ, of a probability model and let M be the most likely value of Θ. The general quadratic expression for Θ is

$$a\Theta^2 + b\Theta + c$$

but we prefer to write it down in the form:

$$\left(-\frac{1}{2}\left(\frac{M-\Theta}{S}\right)^2\right)$$

This expression satisfies the requirement that it takes its maximum value (of zero) when $\Theta = M$.

What we then have to do is to choose the value of S which provides the best approximation to the true log likelihood ratio. Small values of S give quadratic curves with sharp (narrow) peaks and large values of S give quadratic curves with broad peaks.

Eg. For the binomial model we need to choose S so that

$$\left(-\frac{1}{2}\left(\frac{D/N-\pi}{S}\right)^2\right) \quad \cong \quad D\log_e(\pi) + (N\text{-}D)\log_e(1\text{-}\pi).$$

Once M has been found and S chosen, an approximate supported range/confidence interval for Θ is found by solving the equation

$$\left(-\frac{1}{2}\left(\frac{M-\Theta}{S}\right)^2\right) = -1.921$$

to give

$$\Theta = M \pm 1.96S.$$

We shall give formulae for S, without justification, and concentrate on how to use these in practice.

3.1 The risk parameter

The most likely value of π based on D failures and N-D survivors is D/N. To link with tradition we shall also refer to the most likely value of π as P (for proportion). The value of S which gives the best approximation to the log likelihood ratio is

$$S = \sqrt{\frac{P(1-P)}{N}}$$

For the example in which D = 4 and N=10 the value of P is 0.4 and

$$S = \sqrt{\frac{0.4 \times 0.6}{10}} = 0.1549.$$

An approximate supported range for π is given by

$$0.4 \pm 1.96 \times 0.1549 = 0.10 \text{ to } 0.70$$

while the supported range obtained from the true curve lies from 0.15 to 0.70. The true and approximate log likelihood ratio curves are shown in Figure 2. The curve shown as a solid line is the true log likelihood ratio curve, while the broken line indicates the quadratic approximation.

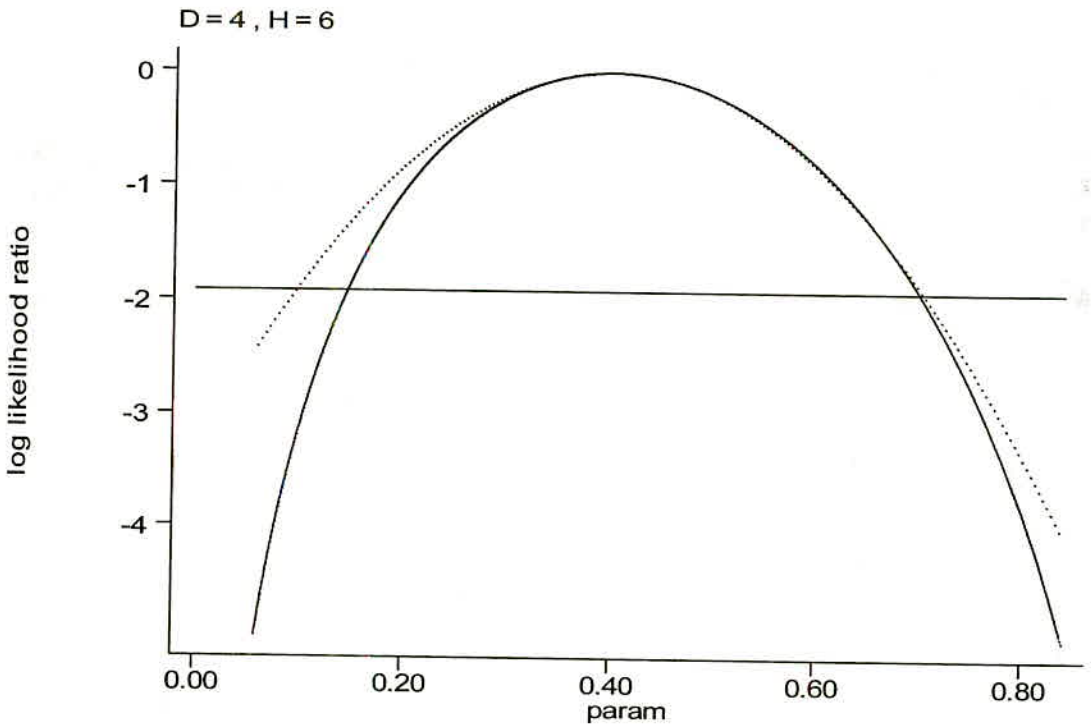


Figure 2. True (solid line) and approximate (dotted line) log likelihood ratios for π

4. Use of approximate likelihoods in tests

4.1 The Wald Test

The quadratic approximation to the log likelihood at the maximum can also be used to calculate the approximate log likelihood ratio for the null value. Tests based on this approximation are called **Wald tests**. The approximate log likelihood ratio for the null value $\Theta = \Theta_0$ is given by

$$-\frac{1}{2}\left(\frac{M-\Theta_0}{S}\right)^2$$

where M is the most likely value of Θ and S is the standard error of the Gaussian approximation.

4.2 The Score Test

Making the quadratic approximation fit well at the most likely value of the parameter is the best way of choosing the approximation for confidence limits since these should be centred around the most likely value. When testing a null hypothesis, an alternative is to find a quadratic approximation which fits as well as possible at the null value of Θ (Figure 3). The test is then called a **score test**. The value of the log likelihood ratio for the null value of the parameter using this approximation is

$$-\frac{1}{2}\frac{U^2}{V}$$

where U is the gradient of the log likelihood at the null value of the parameter, and V is the rate at which this gradient is changing at the null value. The score test is carried out by referring U^2/V to the χ^2 distribution. Since it does not involve the most likely value, the score test is usually much easier to calculate than the Wald test, and many of the commonly used tests in epidemiology (introduced before computers) are score tests. They usually involve calculating the expected number of failures, either overall (E) or in the exposed group (E_1), as shown in Table 1.

Table 1. U and V for some common score tests

Test	Data	E	U	V
$\lambda = \lambda_0$	D, Y	$\lambda_0 Y$	D – E	E
$\pi = \pi_0$	D, N	$N\pi_0$	D – E	$N\pi_0 (1-\pi_0)$
$\lambda_1 = \lambda_0$	D_1, Y_1, D_0, Y_0	$E_1 = DY_1/Y$	$D_1 - E_1$	$DY_1/Y(1-Y_1/Y)$
$\Omega_1 = \Omega_0$	D_1, N_1, D_0, N_0	$E_1 = DN_1/N$	$D_1 - E_1$	$\frac{D(N-D)N_0N_1}{N^2(N-1)}$

One reason why score tests are widely used in epidemiology is that the unstratified score test easily generalises to a stratified version by first calculating U and V separately for each stratum, then summing U and V over strata to obtain the final values of U and V to use in the test. The Mantel-Haenszel procedures are based on score tests.

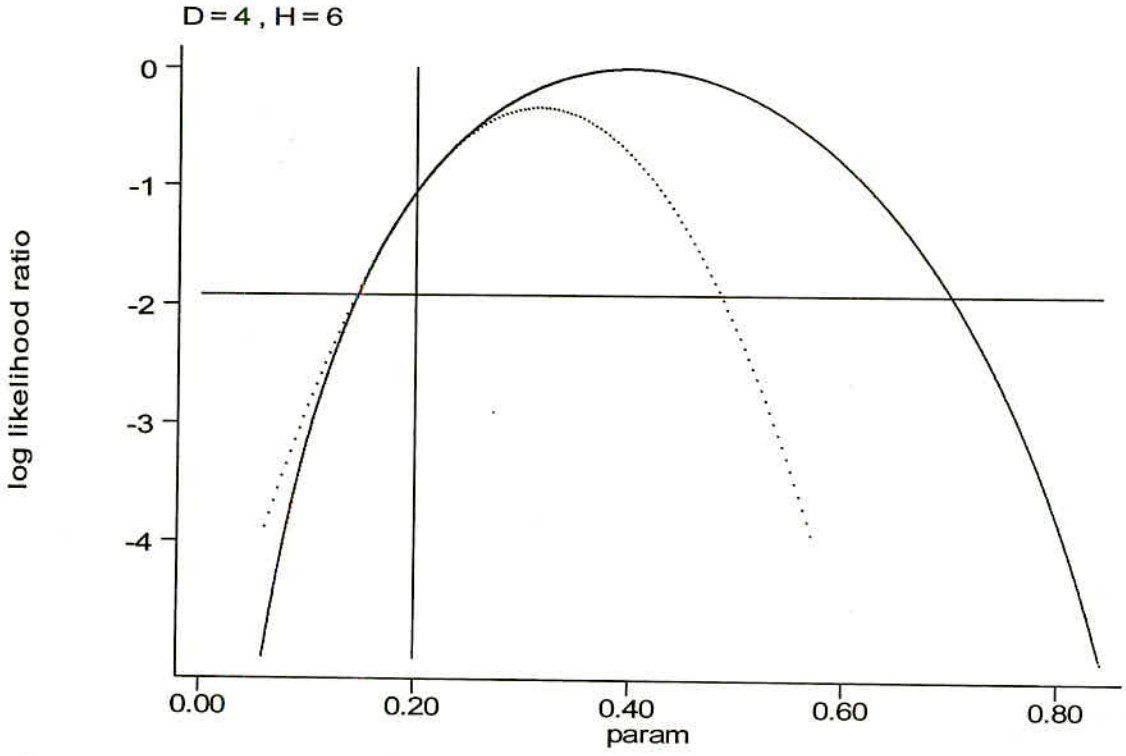


Figure 3. Approximate log likelihood ratio for π (dotted line) where approximation is chosen to fit the true likelihood (solid line) at the null value ($\pi = 0.2$)

4.3. Which test is best?

Three kinds of test have been presented; the **likelihood ratio test**, the **Wald test**, and the **score test**. The second two are based on quadratic approximations to the first, and the likelihood ratio test is itself an approximation when the underlying probability model is not Gaussian (normal). Asymptotically (i.e. when we have an infinite amount of data) they are the same. Fortunately, with a finite but reasonable amount of data all three give very similar answers, unless the null value is very far from the most likely value, in which case the differences between them do not much matter.

5. Transforming the parameter

The risk parameter π must lie between 0 and 1 and the rate parameter λ can take only positive values. Approximate supported ranges for such parameters calculated from a quadratic approximation can sometimes include impossible values. One solution to this problem is to find some function (or transformation) of the parameter which is unrestricted and to first find an approximate supported range for the transformed parameter.

5.1. The log rate parameter

The rate parameter λ can take only positive values, but its logarithm is unrestricted. To calculate an approximate supported range for λ it is better, therefore, to first calculate a range for $\log(\lambda)$, and then to convert this back to a range for λ. The range for $\log(\lambda)$ will always convert back to positive values for λ. To find the range for $\log(\lambda)$ we need a new value of S - that which gives the best

quadratic approximation to the log likelihood ratio curve when plotted against log(λ). When a rate λ is estimated from D failures over Y person-years, this value of S is given by

$$S = \sqrt{(1/D)}.$$

Figure 4 illustrates this new approximation for our example in which D=7 and Y=500 person-years.

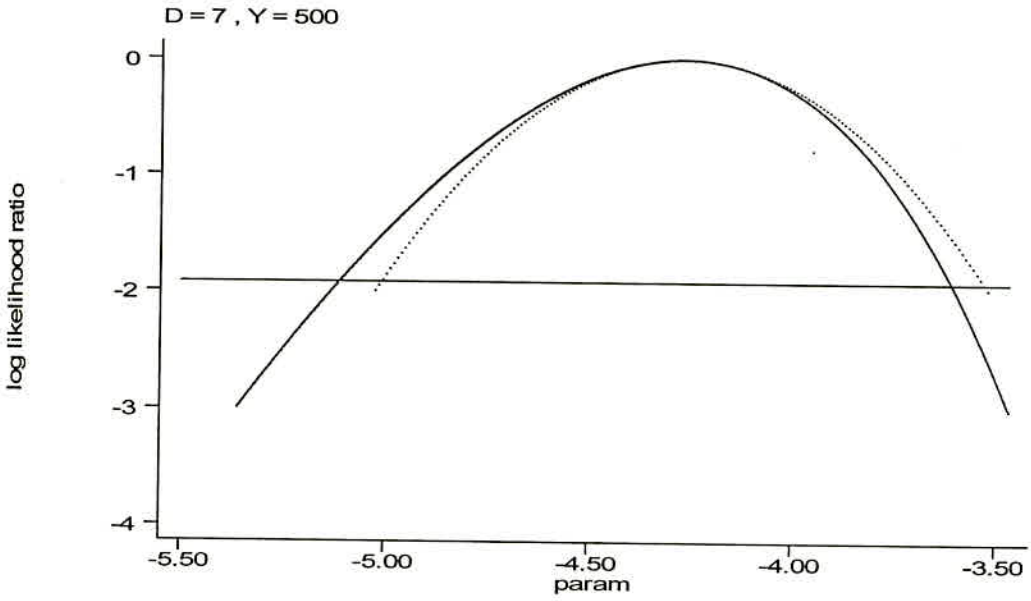


Figure 4. Approximate log likelihood ratio (dotted line) for log(λ)

Here,

$$S = \sqrt{(1/7)} = 0.3780,$$

and an approximate supported range for log(λ) is log(7/500) ± 1.96 × √(1/7), which is from -5.009 to -3.528. The range for λ is from exp(-5.009) = 6.7/1000 to exp (-3.528) = 29.4/1000.

5.2. The log odds parameter

The same thing can be done when calculating a supported range for the risk parameter π based on D failures in N subjects. The value of π is restricted on both sides, by 0 on the left and by 1 on the right. The value of log(π) is still restricted on the right by zero because log(1) = 0, but log(Ω), where Ω is the odds corresponding to π, is not restricted at all. Hence we first find a range for log(Ω) and then convert this back to a range for π. The most likely value of log(Ω) is

$$M = \log(D/(N-D))$$

and the value of S for approximating the log likelihood for log(Ω) is

$$S = \sqrt{\frac{1}{D} + \frac{1}{N-D}}$$

For the example where D = 4 and N-D = 6,

$$S = \sqrt{\frac{1}{4} + \frac{1}{6}} = 0.6455$$

and an approximate supported range for log(Ω) is given by

$$\log(4/6) \pm 1.960 \times 0.6455,$$

that is, from -1.6706 to 0.8597. This is a range for log(Ω) and it is equivalent to a range for Ω from exp(-1.6706) = 0.188 to exp(0.8597) = 2.363. Finally, remembering that $\pi = \Omega/(1+\Omega)$, the range for π is given by

$$\frac{0.188}{1.188} \text{ to } \frac{2.363}{3.363} \text{ which is from 0.16 to 0.70.}$$

6. A table of standard errors

Some of the more commonly used standard errors (S) obtained by approximating the log likelihood are gathered together in Table 2. The formula for the approximate 95% confidence limits is

$$M \pm 1.960\ S.$$

Table 2. Most likely values of parameters with their approximate standard errors

Parameter	Data	Most likely value (M)	Standard error (S)
π	D, N	D/N (= P)	$\sqrt{\frac{P(1-P)}{N}}$
λ	D, Y	D/Y	$\sqrt{D/Y}$
μ	N, $\bar{x}$, σ	$\bar{x}$	$\sigma/\sqrt{N}$
$\log_e(\Omega)$	D, N-D	$\log_e\left(\frac{D}{N-D}\right)$	$\sqrt{\frac{1}{D}+\frac{1}{N-D}}$
$\log_e(\lambda)$	D, Y	$\log_e(D/Y)$	$\sqrt{(1/D)}$
$\pi_1 - \pi_0$	D_1, N_1, D_0, N_0	$P_1 - P_0$	$\sqrt{\frac{P_1(1-P_1)}{N_1}+\frac{P_0(1-P_0)}{N_0}}$
$\lambda_1 - \lambda_0$	D_1, Y_1, D_0, Y_0	$\frac{D_1}{Y_1}-\frac{D_0}{Y_0}$	$\sqrt{\frac{D_1}{Y_1^2}+\frac{D_0}{Y_0^2}}$
$\mu_1 - \mu_0$	$N_1, \bar{x}_1, \sigma_1, N_0, \bar{x}_0, \sigma_0$	$\bar{x}_1 - \bar{x}_0$	$\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_0^2}{N_0}}$
$\log_e(\lambda_1/\lambda_0)$	D_1, Y_1, D_0, Y_0	$\log_e\left(\frac{D_1/Y_1}{D_0/Y_0}\right)$	$\sqrt{\frac{1}{D_1}+\frac{1}{D_0}}$
$\log_e(\Omega_1/\Omega_0)$	D_1, N_1, D_0, N_0	$\log_e\left(\frac{D_1/(N_1-D_1)}{D_0/(N_0-D_0)}\right)$	$\sqrt{\frac{1}{D_1}+\frac{1}{D_0}+\frac{1}{N_1-D_1}+\frac{1}{N_0-D_0}}$

Summary

- likelihood quantifies extent to which data support different values of the parameter of interest
- to obtain point estimate of parameter, choose value which maximises the likelihood
- can also obtain a supported range (confidence interval) by choosing cut-off for the (log) likelihood ratio and perform hypothesis tests
- most likelihoods approximately Gaussian and so most log likelihoods approximately quadratic. We use the quadratic approximation to simplify the calculations
- to avoid getting "impossible" parameter values we usually work with the log rate and the log odds (rather than the rate and the risk)

Appendix The rate parameter (λ)

The likelihood for a rate parameter λ based on D failures and total observation time in years Y is the Poisson likelihood

$$(\lambda)^D \exp(-\lambda Y).$$

The corresponding log likelihood is

$$D\log_e(\lambda) - \lambda Y.$$

Both expressions take their maximum value when $\lambda = D/Y$.

The most likely value of λ based on D cases and Y person years is D/Y and the value of S which gives the best approximation to the log likelihood ratio is

$$S = \frac{\sqrt{D}}{Y}.$$

When D= 7 and Y=500, for example, the most likely value of λ is 0.014 and

$$S = \sqrt{7}/500 = 0.00529.$$

An approximate supported range for λ is therefore

$$0.014 \pm 1.96 \times 0.00529 = 4/1000 \text{ to } 24/1000$$

The true (solid line) and approximate (broken line) log likelihood ratio curves are shown in the figure. The range of support obtained from the true curve spans from 6 to 27 per 1000.

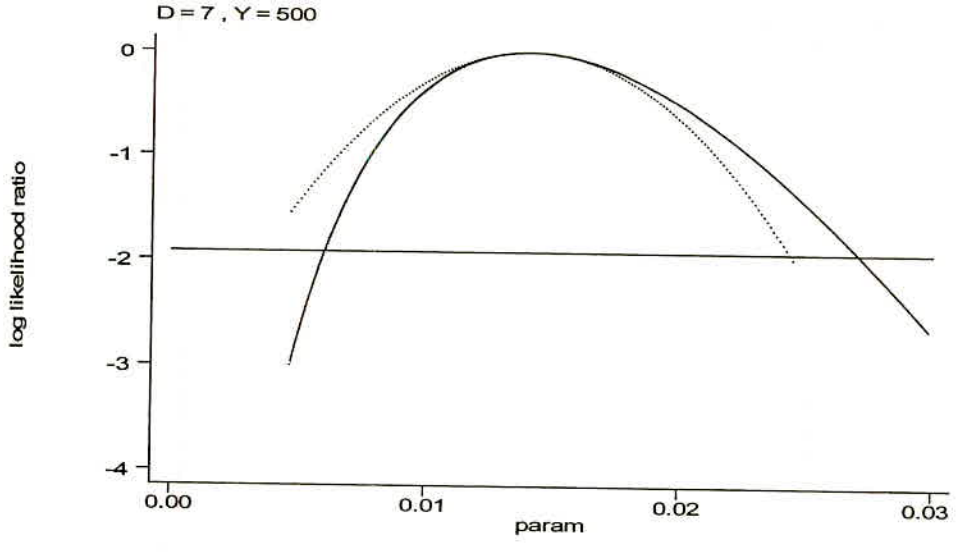


Figure. True (solid line) and approximate (dotted line) log likelihoods for λ.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 8 PRACTICAL

Likelihood

In this practical you will use four commands to explore exact and approximate likelihoods; **`blik`** (binomial likelihood), **`plik`** (Poisson likelihood), **`bloglik`** (binomial log-likelihood), **`ploglik`** (Poisson log-likelihood). These commands are not intended for analysing data - their purpose is only to help you to explore the idea of likelihood.

To obtain help in using these commands type (eg) **`help blik`**.

1. Use **`blik`** to plot the likelihood for π when 4 failures and 6 survivors are observed (**`blik 4 6`**). Do the same for 40 failures and 60 survivors, and for 400 failures and 600 survivors. **Hint:** Use the **`samex`** option to keep scales the same. Note how the 95% interval gets smaller as the total number of subjects increases. This means that the parameter is being estimated more precisely.

2. Try Question 1 with the cut point 0.2585 instead of 0.1465 (option **`cut()`**).

3. Try using **`blik`** with some more extreme splits such as 1:9, 1:99 or 99:1. Notice how the curve is no longer symmetrical and bell-shaped.

4. What happens when the number of failures is zero?

5. Using **`plik`** to plot the likelihood for λ when 7 failures are observed for 500 person years (**`plik 7 500`**). Do the same for 70 failures and 5000 person years, and for 700 failures and 50000 person years. Note how the 95% interval gets smaller as the total number of subjects increases.

6. Try using **`plik`** with a very low number of failures.

7. Out of 25 subjects who are asked to choose between treatments A and B, 15 prefer A and 10 prefer B. Use **`blik`** to plot the likelihood for π, the probability of preferring A (**`blik 15 10`**). What is the most likely value of π? Use the **`null(0.5)`** option to obtain the likelihood ratio for the null value $\pi = 0.5$ (i.e. equally likely to prefer A or B), and use the **`pval`** option to obtain an approximate pvalue for the null value. What are your conclusions?

8. Repeat Question 7 for 40 subjects, of whom 24 prefer A and 16 prefer B. What are your conclusions?

9. Repeat Question 7 for 60 subjects, of whom 36 prefer A and 24 prefer B. What are your conclusions?

Log likelihoods

10. Using **bloglik** to plot the exact and approximate log likelihoods for π (the probability of failure) when there are 4 failures and 6 survivors. Make sure you understand the meaning of all of the output.

11. Use **bloglik** to plot the exact and approximate log likelihoods for π when 40 failures and 60 survivors are observed. What are the 95% limits from the approximation? Notice that the approximation is better than 4 failures and 6 survivors.

12. Use **bloglik** to plot the exact and approximate log likelihoods for π when 400 failures and 600 survivors are observed. Notice that the approximation is now almost perfect.

13. Use **bloglik** to plot the exact and approximate log likelihoods for π when 2 failures and 18 survivors are observed. What are the 95% limits from the approximation? Notice that the approximation is now quite poor because the true log likelihood curve is far from quadratic in shape. Indeed the lower limit is negative, which is not a possible value. Repeat with 20 and 180 and notice that the approximation is now much better.

14. Repeat the first part of Question 13 using the log likelihood plotted against the log odds parameter. What are the 95% limits from the approximation? Notice that when approximating on a log odds scale the confidence limits for π are always between 0 and 1.

15. Use **ploglik** to plot the log likelihood for λ when 7 failures are observed for 500 person-years. What are the 95% limits from this approximation?

16. Repeat Question 15 using the log likelihood plotted against the log rate parameter. What are the 95% limits from the approximation?

17. Use **ploglik** to plot the log likelihood for λ when there is 1 failure for 1000 person years. How good is the approximation? What is the 95% interval for λ? Repeat using log (λ). What is the 95% interval for λ?

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 9

Logistic Regression 1: effect of a single exposure

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

(i) use a logistic model to compare the log odds of disease in two groups i.e. estimate a crude odds ratio for a binary exposure

(ii) perform statistical tests of the null hypothesis – there is no association between the exposure and outcome
- using the Wald test
- using the Likelihood Ratio Test

(iii) use a logistic model to compare the odds of disease for a categorical exposure with 2 or more levels and to estimate crude odds ratios associated with each level..

1. Introduction

In earlier sessions you have seen how to compare rates and odds between exposure groups, both unstratified and stratified. The methods presented can be carried out on a calculator and they are widely used, but they do have limitations. In this and subsequent sessions, we will illustrate the use of a logistic regression model to estimate odds ratios comparing different levels of an exposure variable. The results from logistic regression will be similar to those carried out by hand, but they will be slightly different in some important ways. Topics to be covered in the four logistic regression sessions include:

Logistic Regression 1: effect of a single exposure i.e. using logistic regression to produce crude odds ratios.

Logistic Regression 2: models with more than one variable i.e. using logistic regression to produce odds ratios adjusted for a confounder.

Logistic Regression 3: interaction i.e. using logistic regression to model and test for interactions.

Logistic Regression 4: quantitative exposures i.e. using logistic regression to model and perform tests for trend.

1.1 A reminder of Odds and Odds Ratios (OR).

In earlier lectures we defined the Odds of disease as:

$$\text{Odds} = \frac{\text{number with the disease (D)}}{\text{number without the disease (H)}}$$

Subsequently we saw that the ratio of the Odds between two groups is a useful measure for comparing the frequency of disease in the 2 groups. If one group is 'exposed', and the other 'unexposed', the Odds Ratio for the exposed versus unexposed group is given by:

$$\text{Odds Ratio (OR)} = \frac{\text{Odds in exposed group } (D_1/H_1)}{\text{Odds in unexposed group } (D_0/H_0)}$$

Another way of expressing this relationship is:

Odds in exposed group = (Odds in unexposed group) x (Odds ratio of exposure)

Taking logarithms, the relationship between the exposed group and the unexposed group can now be written:

Log (odds in exposed group) = Log (odds in unexposed) + Log (odds ratio)

In this session we will term "Baseline" to refer to the log odds or odds in the unexposed group, and term "Exposure" to indicate the odds ratio/Log odds ratio. Then we can rewrite the previous equation in the form:

Log odds = Baseline + Exposure.

This is an example of a logistic regression model.

1.2 An introduction to the logistic regression model

Logistic regression models the log odds of disease. The model is written as:

Unexposed person: log odds of disease = Baseline

Exposed person: log odds of disease = Baseline + Exposure

Here "Baseline" represents the log odds of disease in the baseline group, "Exposure" represents the log odds ratio and the model is additive (Baseline + Exposure).

We could rewrite the model using odds instead of log odds as follows:

Unexposed person: odds of disease = Baseline odds

Exposed person: odds of disease = Baseline odds × Exposure OR

Now "Baseline" represents the odds of disease in the baseline group, "Exposure" represents the odds ratio and the model is multiplicative (Baseline × Exposure).

Note that logistic regression does everything on the log odds scale. When it has finished, we (or Stata) can convert back to the odds scale to obtain odds ratios – the things we are really interested in.

1.3 Why model the log odds of disease?

Linear regression is appropriate for response variables like birth weight, that are continuous (quantitative) and have a Normal distribution. It is not appropriate for binary response variables such as diseased/healthy, infected/uninfected, alive/dead. When fitting statistical models for such variables the **log odds** of disease is commonly used as the outcome measure. The reason for modelling the log odds rather than risk or odds is that the log odds can take any value, positive or

negative, whereas risks are constrained to lie between 0 and 1. When using statistical models it is easier to model a quantity which is unconstrained than one which is constrained. This avoids the possibility of predicting impossible values (like risks which are negative or greater than 1) from the model.

Modelling log odds is referred to as **logistic regression**, and the models are referred to as **logistic models**. In this session, we will use logistic regression to model the log odds of disease and hence estimate odds ratios in a cross-sectional study. For convenience, we will speak in terms of 'disease' as the outcome, but everything applies equally to other types of outcome. For the purposes of logistic regression, and these lectures, **all logs are base e.**

2. A logistic model with a binary exposure variable

Example: Data from an onchocerciasis ('river blindness') project which started in 1982 in Sierra Leone will be used to illustrate the methods in this session and later sessions. The project was set up to study epidemiological and clinical aspects of onchocerciasis, in particular the relationship between eye lesions and the prevalence and intensity of microfilariae of *Onchocerciasis volvulus* in skin snips taken from the iliac crest. The study included persons aged ≥5 years who lived in villages in either savannah or forest areas. In this session, we will study the relationship between the outcome, presence or absence of microfilarial infection, and two exposure factors, area of residence (savannah or forest) and age grouped into 4 levels.

2.1 Estimating the odds, log odds and the odds ratio "by hand"

The table shows the prevalence of microfilariae (mf=0 for negative, mf=1 for positive) according to whether the subjects live in the forest (area=1) or savannah (area=0):

Microfil. infection	Area 0	1	Total
0	267	213	480
1	281	541	822
Total	548	754	1302

Before thinking about models, we will calculate the log odds of "disease" (microfilarial infection) and the OR by hand.

Exercise 1: *Fill in the prevalence, odds and log odds of microfilarial infection in the forest and savannah areas.*

	Savannah	Forest	Total
Prevalence			63.1%
Odds			1.712
log odds			0.538

In this example we will take the people living in the Savannah as our baseline group. We can then estimate the Odds Ratio (OR) for the presence of microfilariae for people living in the forest (exposed) compared with people living in the savannah (baseline) as:

$$\frac{541}{213} \div \frac{281}{267} = \frac{2.540}{1.052} = 2.41$$ (This is the cross-product ratio)

Now we will try to summarise these results in a model.

(i) First, using the definition of the odds ratio, we can express the log odds for those in the forest (exposed) in terms of the log odds for those in the savannah (baseline) and the log OR:

OR = (odds in exposed group)/(odds in baseline)

Therefore:

odds in exposed = (odds in baseline) x OR

log (odds in exposed) = log (odds in baseline) + log OR

Substituting the areas in our example gives:

odds in forest = (odds in savannah) x OR (forest compared to savannah)

log (odds in forest) = log (odds in savannah) + log OR(forest vs savannah)

(ii) We can write the second of these two expressions as the logistic regression model:

log odds = Baseline + Area

where Baseline = log (odds in savannah),
Area = log OR for individuals in the forest and 0 individuals in the savannah.

(iii) From the hand calculations we obtained the Odds, odds ratio and log odds of microfilariae infection.

area	**odds**	**log odds of disease**
0=savannah	1.052	0.051
1=forest	1.052 × 2.41 = 2.536	0.051 + 0.881 = 0.932

2.2 Estimating the odds, log odds and the odds ratio by fitting a logistic model in Stata

Recall the table showing the prevalence of microfilariae by area:

Microfil. infection	Area 0	1	Total
0	267	213	480
1	281	541	822
Total	548	754	1302

We can describe the association between area and mf using the logistic model:

$$\text{log odds} = \text{Baseline} + \text{Area}$$

Then we can use Maximum Likelihood to estimate suitable values of Baseline and Area i.e. find the most likely values of these parameters given the data in the above table. The results below have been obtained after fitting this model in STATA. The estimates produced by fitting this simple logistic regression model are identical to the quantities that you calculated by hand.

mf	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
area	.8810211	.1176746	7.487	0.000	.6503832	1.111659
_cons	.051106	.0854637	0.598	0.550	-.1163997	.2186117

(i) The first column notes the names of the outcome variable (`mf`) and the explanatory variable (`area`) in our model. The bottom row labelled `_cons` (meaning constant) refers to the Baseline.

(ii) The second column, labelled `Coef.` (coefficient), gives the Maximum Likelihood estimates of the parameters. The top number, 0.881 is the estimate of the logOR for area while the second number, 0.51, is the estimate of the log odds in the baseline group (savannah). We can substitute the parameter estimates into our model:

$$\text{log odds} = 0.051 + 0.881 \times \texttt{area}$$

where `area`=0 in the savannah and `area`=1 in the forest.

Hence, we can obtain the log odds:

For those in the savannah: $\text{log odds} = 0.051 + 0.881 \times 0 = 0.051$
For those in the forest: $\text{log odds} = 0.051 + 0.881 \times 1 = 0.932$

The output shows the logOR, but we can easily obtain the odds ratio as exp(0.881) = 2.41. **These values for the logOR and OR are exactly the same as those we calculated by hand in Section 2.1.**

(iii) The third column gives the standard errors (`Std. Err.`) of the parameter estimates.

(iv) The fourth column gives z statistics for the parameter estimates, which are used to perform a **Wald test**. The z statistic is calculated as:

z = coefficient/SE

and tests the null hypothesis that the true parameter value (logOR) is 0. The null hypothesis that the logOR is 0, is the same as the null hypothesis that the OR is 1. A p-value (shown in the 5th column) is obtained by comparing the z statistic with a Normal distribution. A 95% confidence interval for the parameter estimate (shown in the final columns) is calculated as:

95% CI for parameter = coefficient ± (1.96 x standard error)

Hence, a 95% confidence interval for the log odds ratio:

95% CI for log OR = log OR ± (1.96 x standard error)

Exercise 2: *Calculate a 95% confidence interval for the log odds ratio for forest vs savannah and confirm that this corresponds to the STATA output. Convert this to a 95% CI for the odds ratio by taking exponentials of this interval.*

The SE of the log Odds Ratio can be calculated from the cells of the 2x2 table (see eg the notes for the 2nd likelihood session) as:

SE(log OR exposed vs unexposed) $= \sqrt{(1/D_1 + 1/(N_1-D_1) + 1/D_0 + 1/(N_0+D_0)}$

SE(log OR comparing forest with savannah) =

95% CI for log odds ratio:

95% CI for odds ratio:

The SE can also be used as an error factor EF=exp(1.96xSE) to calculate a 95% CI for the odds ratio directly:

$$95\%\ \text{CI} = \text{odds ratio} \begin{matrix} \times \\ \div \end{matrix} \text{EF}$$

Thus EF = exp(1.96 x 0.118) = 1.259. Hence, 95% CI = (2.413÷1.259, 2.413x1.259).
= (1.9160, 3.0390)

STATA can provide the Odds Ratios rather than the log odds ratios (logOR). The Odds Ratios are of more familiar to us, and easier to interpret than log(odds ratio)s. In this case the STATA results would be:

mf	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
area	2.413363	.2839914	7.49	0.000	1.916275	3.039397
cons	1.052434	.0899449	0.60	0.550	.8901194	1.244348

Check the odds ratios and 95% CI with your calculations. Check also that the odds ratio is the exponential of the logOR given in the previous STATA output.

This output gives the odds ratio for area 1 (forest) versus area 0 (savannah – the baseline group). We can calculate the odds for each area:

Odds in baseline (savannah) = 1.052

Odds in exposed (Forest) = 1.052 x 2.413 = 2.54

Note: All the estimation is done on the log odds scale before being converted back to the odds scale in the output above. The standard errors are calculated on the log odds scale and so are correct when the output is on the log odds scale. They are **not** correct on the odds scale (above) and the numbers shown should be ignored. They cannot be used directly to obtain confidence intervals. The confidence intervals given above **are** correct, as they are translated directly from the 95% CI on the log odds scale.

3. Testing for association using the Wald test

We have seen that the STATA output includes a **Wald test** for each parameter. The null hypothesis for this test is that the true parameter value is 0. The test statistic (z) is obtained by dividing the parameter estimate by its SE and comparing it with a Normal distribution.

The Wald test for `area` assesses the null hypothesis that the true log OR=0 (i.e. that the true odds ratio is 1) versus the alternative that the true log OR is not 0. The Wald test for the association between microfilarial infection and area is given by:

$$z = \log(OR)/SE(\log OR) = 0.881/0.118 = 7.487$$

As shown in the STATA output, the corresponding p-value is small ($p<0.001$), indicating strong evidence against the null hypothesis of no association between microfilarial infection and area.

The Wald test for the baseline odds is not usually useful, and should be ignored. This tests the null hypothesis that the baseline odds are 1 (ie 50% have the disease, and 50% do not).

4. Testing for association using the Likelihood ratio test

Recall from the likelihood sessions that the log likelihood in terms of the odds parameter Ω is:

$$D\log(\Omega) - N\log(1+\Omega)$$

where D is the number of persons with the disease, N is the total number of persons and Ω is the odds of disease. The **most likely** value of Ω is given by the **observed odds**, D/(N-D).

For any statistical test, we first need to state our null hypothesis. In this case the null hypothesis is that the odds of microfilarial infection are the same in the forest and in the savannah. In other words the odds and the **log likelihood** of the odds are calculated from the overall numbers infected and not infected:

Number NOT infected	= 480	
Number infected (D)	= 822	
Total number (N)	= 1302	
Odds (Ω)	= 822/480	= 1.713

Under the null hypothesis the Log Likelihood can be worked out as: $D \log(\Omega) + N \log(1+\Omega)$

Worked Practical: *Calculate the log likelihood for the null model.*

$$L_0 = 822 \log(1.713) - 1302 \log(2.713)$$

$$= -857.0$$

In STATA we can obtain the logistic regression output for the null hypothesis, and obtain the Log Likelihood under the null hypothesis.

```
Log likelihood = -857.02925
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |      1.7125   .0983736     9.36   0.000     1.530149    1.916582
------------------------------------------------------------------------------
```

The **Likelihood Ratio Test** (LRT) is based on the Likelihood Ratio Statistic (LRS):

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

where L_1 is the maximised log likelihood under the alternative hypotheis, in this case that there are different odds of disease in each group, and L_0 is the log likelihood under the null hypothesis calculated above.

For the comparison of the two areas, forest and savannah, in our example, we can calculate the log likelihood under from the odds in the 2 areas. This will be the sum of the 2 log likelihoods:

The log likelihood for the odds parameter Ω_0 in the savannah is:

$$D_0\log(\Omega_0) - N_0\log(1+\Omega_0)$$

and the log likelihood for the odds parameter Ω_1 in the forest is:

$$D_1\log(\Omega_1) - N_1\log(1+\Omega_1)$$

where D_0 , N_0 , Ω_0 , D_1 , N_1 , Ω_1 are shown in the following table:

Microfil. infection	Area 0	Area 1	Total
0	267	213	480
1	281 (D_0)	541 (D_1)	822
Total	548 (N_0)	754 (N_1)	1302
odds	1.052 (Ω_0)	2.540 (Ω_1)	

Exercise 3: *Calculate the values of L_1 and hence the likelihood ratio statistic $2(L_1-L_0)$, using the value of L_0 obtained in the previous exercise.*

L_1 = [281 log(1.052) - 548 log(1+1.052)] + [541 log(2.540) - 754 log(1+2.540)]

L_1 =

LRS = $2(L_1 - L_0)$ =

Under the null hypothesis, the LRS is distributed as χ^2 on 1 d.f. The test has one degree of freedom in this instance because we are testing one parameter in this example – the log(odds ratio) for area.

The corresponding P value is again small ($p<0.001$). We will see in the practical how to perform a LRT in STATA.

Since a χ^2 statistic on 1 df is just the square of a z statistic, the LRS should be approximately equal to the square of the z statistic calculated for the Wald test. In this instance, $z^2 = 7.487^2 = 56.05$ and LRS = 57.025. These results are also very close to that obtained using the usual chi-squared test for a 2 x 2 table which gives $\chi^2 = 56.3$.

Summary so far

We want to estimate the strength of the association between area and microfilarial infection. We can perform a logistic regression analysis in STATA (or another package):

mf	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
area	.8810211	.1176746	7.487	0.000	.6503832	1.111659
_cons	.051106	.0854637	0.598	0.550	-.1163997	.2186117

We can use the output to obtain the following:

(i) Estimates of the log odds or odds of disease

For those in the savannah:	log odds	= baseline	= 0.051 + 0.881x0	= 0.051
	odds	= exp(0.051)	=1.051	
For those in the forest:	log odds	= baseline + log OR	= 0.051 + 0.881x1	= 0.932
	odds	= exp(0.932)	= 2.541	

(ii) OR and 95% CI

log OR	= 0.881,	
OR	= exp(0.881)	= 2.41
95% CI	= exp(logOR ± 1.96×Std Err)	= 1.91 - 3.04

(iii) Wald test

To test the null hypothesis that the true odds ratio=1 (logOR=0) versus the alternative hypothesis that the true odds ratio is not 1:

$$z = (\log OR)/SE = 0.881/0.118 = 7.487, \quad p<0.001$$

5. Logistic model for the comparison of more than two groups

We can also use logistic regression to examine the association between the outcome and an explanatory variable with more than 2 levels. Here, all odds ratios are calculated relative to the **baseline** group (by default the group with the lowest coded value of the explanatory variable).

Example: The table below shows the prevalence of microfilariae by age group where agegroup is coded as 0 (5-9 years), 1 (10-19 years), 2 (20-39 years), 3 (≥40 years):

microfil. infection	Age group 0	1	2	3	Total
0	156	119	125	80	480
1	46	99	299	378	822
Total	202	218	424	458	1302

First, we will estimate the log odds and the odds ratios "by hand".

Exercise 4 : *Complete this table by calculating the odds, log odds, odds ratios (compared to age group 0) and log odds ratios as required.*

	Age group (yrs) 5-9	10-19	20-39	≥40
Odds	0.29	0.83	2.392	
Odds Ratio	1.00	2.82		16.03
log odds	-1.221		0.872	
log OR	0	1.037		2.774

Note that you can calculate the log odds ratio for age group 10-19 compared to age group 5-9 either as:

1.037 = log(2.82) (the log OR), or

1.037 = (-0.184) - (-1.221) (the difference in the log odds).

Check that this works for the results you calculated for the other two age groups.

We can describe the association between age group and mf infection using the logistic model:

log odds = Baseline + Agegrp

where Baseline is the logodds in the lowest age group (ge group 5-9)
Agegrp is the logOR for each level of age group relative to age group 5-9 (three non-zero logORs)

Fitting this model in STATA produces the following output:

mf	Coef.	Std. Err.	z	P> |z|	[95% Conf.	Interval]
_Iagegr_1	1.037211	.2159948	4.802	0.000	.6138689	1.460553
_Iagegr_2	2.093344	.1987306	10.534	0.000	1.70384	2.482849
_Iagegr_3	2.774082	.2080735	13.332	0.000	2.366266	3.181899
_cons	-1.221215	.1677778	-7.279	0.000	-1.550053	-.8923762

(i) The first column indicates the name of the outcome variable (`mf`) and explanatory variables (`_Iagegr_1, _Iagegr_2, _Iagegr_3`) in our model.
(ii) The second column, labelled `Coef.` (coefficient), gives the Maximum Likelihood estimates of the parameters. The estimate of the logOR for `agegrp` = 1 is 1.037 (`_Iagegr_1`), of the logOR for `agegrp` = 2 is 2.093, of the logOR for `agegrp` = 3 is 2.774. Note that there are 3 estimates because age group has 4 levels, one of which is the baseline group and which is not shown in the output.

The estimate of the log odds in the baseline group (age group 5-9) is -1.221. We can substitute the parameter estimates into our model and obtain the log odds and the odds ratios:

$$\text{log odds} = -1.221 + 1.037 \times \texttt{\_Iagegr\_1} + 2.093 \times \texttt{\_Iagegr\_2} + 2.774 \times \texttt{\_Iagegr\_3}$$

`_Iagegr_1, _Iagegr_2, _Iagegr_3` are **indicator variable** created by STATA for each non-baseline value of the categorical age group for the purposes of the analysis. Indicator variables take only the values 0 and 1. STATA labels these as `Ivarname_#` where # is the level of the variable (for details, see practical session).

`_Iagegr_1` is an indicator variable that equals 1 for age group 10-19 & equals 0 otherwise,
`_Iagegr_2` is an indicator variable that equals 1 for age group 20-39 & equals 0 otherwise,
`_Iagegr_3` is an indicator variable that equals 1 for age group 40+ & equals 0 otherwise.

For those in agegroup 5-9: log odds = -1.221 + 1.037x0 + 2.093x0 + 2.774x0 = -1.221
Odds = 0.29

For those in agegroup 10-19: log odds = -1.221 + 1.037x1 + 2.093x0 + 2.774x0 = -0.184
Odds = 0.294 x 2.82 = 0.83

For those in agegroup 20-39: log odds = -1.221 + 1.037x0 + 2.093x1 + 2.774x0 = 0.872
Odds = 0.294 x 8.11 = 2.38

For those in agegroup 40+: log odds = -1.221 + 1.037x0 + 2.093x0 + 2.774x1 = 1.553
Odds = 0.294 x 16.02 = 4.71

These values of the odds, log odds and the odds ratios are the same (apart from rounding errors) as those we calculated by hand in Exercise 4. The remaining columns gives the standard errors, z statistics (for the **Wald test**), the Wald test p-value, and a 95% CI for the parameter. Note that there is a separate Wald test for each agegrp (each one assesses the null hypothesis that a particular log OR=0).

The same logistic regression command asking for the Odds Ratios gives the following output:

```
Logistic regression                               Number of obs   =       1302
                                                  LR chi2(3)      =     258.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -727.83149                       Pseudo R2       =     0.1508
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagegrp_1 |   2.821337   .6093937     4.80   0.000     1.847566    4.30834
  _Iagegrp_2 |      8.112   1.612101    10.53   0.000     5.495007   11.97533
  _Iagegrp_3 |   16.02391   3.334121    13.33   0.000     10.65756   24.09236
        cons |   .2948718   .0494729    -7.28   0.000     .2122368   .4096809
------------------------------------------------------------------------------
```

Note with the model with 4 agegroups is compared to model under the null hypothesis (given earlier in the session). The LR test statistic can be obtained as $2(L_1-L_0)$, and compared to the chi-squared distribution on 3 d.f.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 9: PRACTICAL

Use of STATA for simple logistic regression analyses

By the end of this practical students will be able to:

(i) Use the **tabodds** and **mhodds** commands for estimating odds and odds ratios in a cross-sectional study.
(ii) use the **logistic** and **logit** commands to compare 2 or more groups in a crude analysis.
(iii) use the **xi** command in front of the **logistic** command to analyse exposures with more than 2 categories, with **i.** in front of the exposure variable name
(iv) use the **lrtest** command to conduct a likelihood ratio test.

1. Ensure that the course data files have been copied into your user area, then start STATA.

2. The data are in the dataset onchall. Type `use onchall` to use the dataset. 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).**

3. The outcome variable is `mf`, which is coded as 1 for an individual with microfilarial infection, 0 for an uninfected individual. Use the `tab`, `tabodds` and `mhodds` to examine the association between `mf` and `area`.

4. To use logistic regression to examine the association between `mf` and `area`, type

```
logit mf area
```

Check that the last part of the output corresponds to that in the lecture notes. The first part of the output is:

```
Iteration 0:      Log Likelihood = -857.02925
Iteration 1:      Log Likelihood = -828.59538
Iteration 2:      Log Likelihood = -828.51659
Iteration 3:      Log Likelihood = -828.51659

Logit Estimate                          Number of obs     =   1302     (i)
                                        chi2 (1)          =  57.03    (ii)
                                        Prob > chi2       =  0.0000 (iii)
Log Likelihood = -828.51659             Pseudo R2         =   0.0333 (iv)
```

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.

5. Check that the log likelihood for the model agrees with what you calculated in the lecture. The statistics in the column on the right of the previous STATA output are:

   (i) the number of observations
   (ii) A likelihood ratio χ^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. Check that this agrees with the likelihood ratio statistic which you calculated in the lecture.
   (iii) the p-value for the likelihood ratio test.
   (iv) A 'goodness of fit' statistic (we are not interested in this for the purposes of this course).

6. So far, all the output has been on the log scale. This was needed in order to explain how confidence intervals and Wald tests (z-tests) are derived. In fact, STATA allows us to derive estimates on the odds ratio scale, which is much more convenient for reporting results. Type:

```
logistic mf area
```

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

   (i) The baseline term (`_cons`) is omitted when the `logistic` command is used.
   (ii) The standard error for the odds ratio is only approximate and should be ignored.
   (iii) The z statistic (for the Wald test) is identical to that when the `logit` command is used, and is derived using the standard error for the log odds ratio.
   (iv) The confidence interval is also derived using the standard error for the log odds ratio, as shown in the lecture. Hence, the confidence interval is correct.

Another way to get exactly the same output is

```
logit mf area, or
```

(the `or` option stands for odds ratio).

7. The output for this command shows the Odds Ratios for the comparison of each level of the exposure against the baseline, but it does not give the Odds of microfilariae infection in the baseline group. To get STATA to show the baseline odds we need to create a new variable and substitute it for the constant. Type:

```
gen cons = 1
logit mf area cons, noconstant or
```

The option `noconstant` in the logit command suppresses the constant term in the model. We have included an extra term in the model to make up for this, this extra

term gives us the Odds of infection in the baseline group (savannah). This can only be done using the logit command (not the logistic command), and the odds ratio of the baseline is given if the extra term is a constant equal to one.
The option `or` tells STATA to give the Odds Ratios in the output.

8. Use the `tab` command to examine the association between age group and microfilarial infection. Are column or row percentages more appropriate in your table?

9. As explained in the lecture, we need to use indicator variables to examine the association between agegroup and microfilarial infection. To do this, type

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

We need to type `xi:` before the `logit` (or `logistic`) command to tell STATA that we will be using categorical variables (`xi` stands for expand indicator). To tell STATA that `agegrp` is a categorical variable, we will simply type `i.` in front of the variable name.

10. To get the output on the odds ratio scale, type:

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

Check that the odds ratios are the same as those which you calculated in the lecture, and **check that you understand all the output and how it is derived.**

Note that there are three odds ratios each of which refers to the same baseline group (those aged 5-9). The odds ratio is 2.82 for those aged 10-19 compared to those aged 5-9, 8.11 for those aged 20-39 compared to those aged 5-9, and 16.02 for those aged ≥40 compared to those aged 5-9. There are **three Wald test p-values** (one for each odds ratio) which test whether each odds ratio is significantly different to one. 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 **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).

11. Because `area` is already coded as 1 and 0, it makes no difference if it is used as a categorical variable. Check this by typing the two commands:

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

12. We will now describe the general method for deriving likelihood ratio statistics in STATA. To do this, we need to get STATA to compare the log likelihood from the model with agegrp (L_1) and the log likelihood from the model without agegrp (L_0). L_1 is obtained first and saved in STATA in something called "0". Then L_0 is obtained and compared with L_1. Type the following four commands:

`xi: logistic mf i.agegrp` (fit the model with agegrp)
`estimates store A` (save L_1 = the log likelihood from this model in something called "A")
`logistic mf` (fit the null hypothesis model i.e. omit `agegrp`)
`estimates store B` (save L_0 in "B")
`lrtest B A` (compare L_0 with L_1, and give the likelihood ratio test, on appropriate d.f.)

Note that in this simple example, the LRS is the same as the model X2 statistic which is given in the top right hand corner of the output (explained in 5 ii)).

13. Examine the association between `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?

14. Fit a model with both `area` and `agegrp` as explanatory variables, by typing:

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

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

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 10

Logistic regression 2: models with more than one variable

Objectives

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

(i) use logistic regression to examine the association between an exposure and disease adjusting for confounders, assuming no effect modification
(ii) explain the implications of assuming no effect modification/interaction
(iii) use the Likelihood Ratio Test to test the association between an exposure and disease after adjusting for confounders
(iv) explain why logistic regression can be used for unmatched case-control studies.

1. Introduction

In earlier sessions, we saw how to control for a confounding variable by dividing the sample into strata defined by levels of the confounder. We estimated the odds (or rate) ratio in each stratum and combined the stratum-specific estimates into an overall summary odds (rate) ratio using the Mantel-Haenszel method (by calculating a weighted average of the stratum-specific estimates). This approach assumes that the true odds (rate) ratio for exposed versus unexposed individuals is constant over the different strata. In this session we shall see how to perform similar analyses using logistic regression.

2. Adjusting for confounding using the Mantel-Haenszel method

The numbers of individuals with and without microfilarial infection according to area and age group are shown below.

	area							
mf	0 (sava)	1 (forest)	0	1	0	1	0	1
1=yes	16	30	22	77	123	176	120	258
0=no	77	79	50	69	85	40	55	25
	agegrp =0 (5-9)		**agegrp** =1 (10-19)		**agegrp** =2 (20-39)		**agegrp** =3 (40+)	

Exercise 1

a). From the raw data it is possible to calculate the odds of infection for each combination of area and age group. For example, for those in **agegrp** 0 who live in the savannah (**area** = 0), the odds of infection are 16/77=0.208. In the table below, fill in the missing odds of infection and odds ratios for area in each age group.

	odds		
agegrp	**area** = 0	**area** = 1	odds ratio
0	0.208		
1	0.440		
2	1.447		
3	2.182		

```
. mhodds mf area, by(agegrp)

-------------------------------------------------------------------------------
   agegrp | Odds Ratio      chi2(1)       P>chi2       [95% Conf. Interval]
----------+--------------------------------------------------------------------
        0 |   1.827532         3.02       0.0821         0.91662    3.64368
        1 |   2.536232         9.53       0.0020         1.37466    4.67931
        2 |   3.040650        25.39       0.0000         1.92880    4.79343
        3 |   4.730000        38.21       0.0000         2.74594    8.14764
-------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for agegrp
    ------------------------------------------------------------
     Odds Ratio    chi2(1)       P>chi2       [95% Conf. Interval]
    ------------------------------------------------------------
       3.039445      69.89       0.0000         2.310399   3.998542
    ------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(3)    =     5.05
                                     Pr>chi2    =   0.1680
```

Performing a Mantel-Haenszel analysis of these data, we obtain stratum specific estimates of the odds ratio for the effect of area in each age group (1.83, 2.54, 3.04, 4.73), a summary, weighted average of these stratum specific estimates (3.04), and a test of the null hypothesis that the true odds ratio is the same in each age group (i.e. no effect modification/interaction) versus the alternative that the odds ratio varies across strata.

What do we conclude from this output? Although the odds ratio for the effect of area appears to increase with age, the statistical test for interaction suggests that the observed variation may be no greater than one might expect to observe by chance. If we are prepared to assume no effect modification then it is reasonable to compute a summary, age-adjusted OR (3.04) which we can then compare with the crude odds ratio that we obtained in the previous session (2.41). The crude and adjusted ORs differ somewhat, suggesting some confounding with age, with a failure to take age into account in the analysis leading to an underestimate of the effect of area. We can also use the Mantel-Haenszel approach to obtain estimates of the effect of age adjusted for area.

```
. mhodds mf agegrp area, c(1,0)
Mantel-Haenszel estimate of the odds ratio
Comparing agegrp==1 vs. agegrp==0, controlling for area
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
2.631592	19.89	0.0000	1.691526	4.094101

```
mhodds mf agegrp area, c(2,0)
Mantel-Haenszel estimate of the odds ratio
Comparing agegrp==2 vs. agegrp==0, controlling for area
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
9.042803	132.58	0.0000	5.735696	14.256731

```
mhodds mf agegrp area, c(3,0)
Mantel-Haenszel estimate of the odds ratio
Comparing agegrp==3 vs. agegrp==0, controlling for area
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
16.639417	215.78	0.0000	9.976373	27.752590

For the sake of brevity we have not shown here the stratum-specific estimates or tests for effect modification/interaction, but we should usually look at them to determine whether summary estimates are appropriate. Comparing the estimates for the effect of age group adjusted for age (2.63, 9.04, 16.64) with the crude estimates obtained in the previous session (2.82, 8.11, 16.02), the differences are not large suggesting only limited confounding of age group by area.

3. Adjusting for confounding using logistic regression

At the end of the last practical session you were asked to fit a logistic regression model including terms for both area and age group.

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

mf	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iarea_1	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
constant	.1473754	.0289699	-9.74	0.000	.1002542	.2166443

How should we interpret the parameter estimates above?

We need to remember that logistic regression models the log odds. Thus the above model specifies that:

log(Odds of infection) = Baseline + Area + Age group

where "Baseline" represents the log(odds of infection) in the baseline group (`area` = 0 and `agegrp` = 0),
"Area" and "Age group" represent the log(Odds ratios) associated with `area` and `agegrp`.

On the odds/odds ratio scale this becomes:

Odds of infection = Baseline × Area × Age group

where "Baseline" now represents the odds of infection in the baseline group and "Area" and "Age group" now represent the odds ratios associated with `area` and `agegrp`.

Component of model	**Parameter name in output**	**Estimate**	**Interpretation**
Baseline	`constant`	0.147	Odds in baseline group
Area	not shown (`area` = 0)	1.0 (fixed)	
	`_Iarea_1` (`area` = 1)	3.083	Odds ratio, `area` 1 vs `area` 0
Agegrp	not shown (`agegrp` = 0)	1.0 (fixed)	
	`_Iagegrp_1` (`agegrp` = 1)	2.599	Odds ratio, `agegrp` 1 vs `agegrp` 0
	`_Iagegrp_1` (`agegrp` = 2)	9.765	Odds ratio, `agegrp` 1 vs `agegrp` 0
	`_Iagegrp_1` (`agegrp` = 3)	17.642	Odds ratio, `agegrp` 1 vs `agegrp` 0

We can use these parameter estimates to calculate the fitted odds for each combination of area and age group. For example, the fitted odds for individuals in area 1, age group 1, are given by:

Odds of infection = Baseline × Area × Age group

= 0.147 × 3.083 × 2.599

= 1.178

Exercise 2

Fill in the empty cells in the table below with the fitted odds from the model:

Odds of infection = Baseline × Area × Age group

Agegrp	Area: 0 (savannah)	Area: 1 (forest)
0 (5–9)	0.147	
1 (10-19)		1.178
2 (20-39)		
3 (40+)		

What is the odds ratio from the model for the effect of area in age group 0? Age group 1? Age group 2? Age group 3?

In Exercise 1 you calculated the observed odds and odds ratios as:

Agegrp	Area: 0	Area: 1	Odds ratio
0	0.208	0.380	1.828
1	0.440	1.116	2.536
2	1.447	4.400	3.041
3	2.182	10.32	4.730

When we calculate the observed odds and odds ratios as above, we make no assumption about how the effects of area and age group combine. In the model that we have fitted, the assumption has been made that the effect of area (i.e. the odds ratio) is the same in each age group; i.e. we have assumed that the effect of area of area in not modified (changed) by age group - that age group is not an effect modifier for area/does not interact with area. The parameter estimate for the effect of area that we obtained from the model (odds ratio = 3.083) represents the (summary) odds ratio for the effect of area, adjusted for any confounding effect of age group. Notice that the estimate we obtain from logistic regression (3.08) is very close to the summary Mantel-Haenszel estimate we obtained previously (3.04). Both of these estimates can be thought of as weighted averages of the stratum-specific odds ratios. They are not identical because they use different sets of weights. The age group parameters are interpreted in a similar fashion: they are the estimated odds ratios for the effect of age group adjusted for any confounding effects of area. They too are very similar to the analogous estimates obtained using the Mantel-Haenszel approach.

Warning note: our logistic regression analysis has produced simultaneously a summary estimate of the effect of area adjusted for age group, and summary estimates of the effect of age group adjusted for area. It has done this making the assumption that there is no effect modification/interaction between the effects of area and age group. While the output from the

Mantel-Haenszel analyses provides us with reminders to think about effect modification/interaction (in the form of stratum-specific estimates and a test for effect modification), the output from the logistic regression analysis does not. It is very easy when using logistic regression analysis to make the assumption of no effect modification without realising it. In the next session we shall see how to investigate possible effect modification using logistic regression models.

4. Hypothesis tests in logistic models with more than one variable

When we fitted the model including the effects of area and age group we obtained the following output:

mf	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iarea_1	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
constant	.1473754	.0289699	-9.74	0.000	.1002542	.2166443

Suppose that we wanted to perform a statistical test of the null hypothesis:

> **H_0 :** after taking account of the effects of age group there is no association between area and odds of mf infection

versus the alternative hypothesis:

> **H_1 :** after taking account of the effects of age group there is an association between area and odds of mf infection

We could use the z-statistic (8.18) and p-value ($0.000 \Rightarrow p<0.001$) provided with the output to conclude that there is strong evidence against the null hypothesis. In doing this we would be performing a Wald test (based on the quadratic approximation to the log likelihood at its maximum).

Now suppose that we wanted to perform a statistical test of the null hypothesis:

> **H_0 :** after taking account of the effect of area there is no association between age group and odds of mf infection

versus the alternative hypothesis:

> **H_1 :** after taking account of the effect of area there is an association between age group and odds of mf infection

The output provides us with three p-values, one for each of the age group parameters. Given that all three p-values are very small, and the confidence intervals do not all overlap, we might feel safe in concluding that, as before, there is strong evidence against the null hypothesis. However, what should we do when things are not so clear cut.

Another way of stating H_0 is as follows:

after taking account of the effect of area,
the odds ratio for age group 1 versus age group 0 is 1.0 **and**
the odds ratio for age group 2 versus age group 0 is 1.0 **and**
the odds ratio for age group 3 versus age group 0 is 1.0.

We can test simultaneously that all three odds ratios are equal to 1.0 by performing a likelihood ratio test (LRT). Remember that a LRT is performed by comparing:

L_1 =the log likelihood when the parameters of interest take their most likely value, and

L_0=the log likelihood when the parameters of interest are zero (the null value)

The test is based on comparing the value of the Likelihood Ratio Statistic (LRS) = $2 \times (L_1 - L_0)$ with the chi-squared distribution. How do we do this in practice?

Step 1. Obtain the value of L_1 by fitting a model which allows the odds ratios for age group to takes their most likely values after taking account of area (i.e. fit a model with both area and age group in it).

Step 2. Save the log likelihood (L_1)

Step 3. Obtain the value of L_0. This requires us to fit a model in which age group is assumed to have no effect (all three odds ratios are 1 or, equivalently, all three log(odds ratios) are 0) – i.e. a model without age group, but with area, in it.

Step 4. Save the log likelihood (L_0).

Step 5. Compare L_1 and L_0.

```
xi:logistic mf i.agegrp i.area          /* STEP 1 */

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]
-------------+----------------------------------------------------------------
  _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
    _Iarea_1 |   3.083224    .424372     8.18   0.000     2.354217    4.037975
------------------------------------------------------------------------------
```

```
estimates store A                                   /* STEP 2 */

xi:logistic mf i.area                               /* STEP 3 */

Logistic regression                                 Number of obs   =       1302
                                                    LR chi2(1)      =      57.03
                                                    Prob > chi2     =     0.0000
Log likelihood = -828.51659                         Pseudo R2       =     0.0333

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.413363   .2839914     7.49   0.000     1.916275    3.039397
------------------------------------------------------------------------------

estimates store B                                   /* STEP 4 */

lrtest A B                                          /* STEP 5 */

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

$L_1 = -692.40733$ and $L_0 = -828.51659$

$$\Rightarrow 2(L_1 - L_0) = 2 \times (-692.407 + 828.517) = 272.22$$

We are testing three age group parameters tested simultaneously (all odds ratios =1), so the two models we have fitted differ by three parameters (`_Iagegrp_1`, `_Iagegrp_2`, `_Iagegrp_3`). Hence the LRT has 3 degrees of freedom; i.e. we compare the LRS with the chi-squared distribution with 3 degrees of freedom.

The result of the LRT ($p<0.0001$) tells us that there is strong evidence against the null hypothesis; i.e. there is strong evidence that, after taking account of area, there is an association between age group and odds of infection. I.e., there is strong evidence that, even after taking account of area, age group is a "risk factor" for mf infection.

Note that the LRT tells us whether there is evidence that a variable (age group) is a risk factor for infection. It is not a test for whether age group is a confounder. No statistical test for confounding exists. To be a confounder, the variable (age group) must be a risk factor for infection **and** associated with exposure of interest (area). The LRT we have performed only examines the risk factor part. In order to assess whether a variable is a confounder, we compare the crude and adjusted odds ratios.

Further note: the output from fitting a logistic regression model presents the results of a LR test. E.g. from the model with both area and age group:

```
LR chi2(4)      =     329.24
Prob > chi2     =     0.0000
```

Notice that this test has 4 degrees of freedom. This is because it is testing the null hypothesis that there is no association between odds of mf infection and either age group (3 parameters) or area (1 parameter). The result provides strong evidence against the null hypothesis. However, this is not a very useful test to perform because we are testing two things (age group and area) at once. We cannot tell whether this result is indicating evidence of an association of mf infection with age group or with area or both. For models containing more than one variable, the LR test result presented at the top of the output is not generally of much use.

5. Logistic regression in case-control studies

Logistic regression models the log(odds) and hence odds ratios. We have introduced logistic regression in the context of a cross sectional study in which we can estimate the odds in each exposure group. In a case–control study we cannot estimate the odds of infection/disease for reasons discussed in a previous session, yet we can estimate the odds ratio. Can we use logistic regression to analyse data from case-control studies?

Suppose that in a population we have disease associated with a single exposure with just two levels (coded as 0 and 1) as follows.

	Exposed	Unexposed	Total
Diseased	D_1	D_0	D
Healthy	H_1	H_0	H
Total	N_1	N_0	N

Now suppose that we conduct a case-control study in this population. Cases and controls are recruited with different sampling fractions, S_D for cases and S_H for controls. Then the expected result of the study are as shown below:

	Exposed	Unexposed	Total
Cases	$S_D \times D_1$	$S_D \times D_0$	$S_D \times D$
Controls	$S_H \times H_1$	$S_H \times H_0$	$S_H \times H$

We saw in an earlier session that although we cannot estimate the absolute risk or odds of disease from a case control study, we can estimate the odds ratio using simple methods. If we naively attempted to estimate the baseline odds we would be estimating the quantity:

$$(S_D \times D_0)/(S_H \times H_0) = (S_D/S_H) \times (D_0/H_0)$$

$$= (S_D/S_H) \times \text{True Baseline Odds}$$

So what happens when we fit a logistic regression model to case control data is that we model

"Odds" in exposed = $[(S_D/S_H) \times$ True Baseline Odds$] \times$ Exposure

The **constant** term estimated by logistic regression no longer estimates the true baseline odds, but estimates ($S_D/S_H \times$ the baseline odds). Since we don't usually know what S_D and S_H are, this means that the **constant** term in the model is usually un-interpretable. However odds ratio estimate is correct.

Example

Consider a population of 30,000 initially disease-free individuals observed for 10 years after which time 50 subjects were diseased. The results of the study might be tabulated as follows:

		Exposure		
		+	-	Total
Disease	+	30	20	50
	-	9970	19980	29950
		10000	20000	30000

The true odds ratio is $(30 \times 19980)/(20 \times 9970) = 3.006$.

The odds in the baseline group = 20/19980 = 0.001001

Fitting a logistic regression model to these data produces the expected results:

```
------------------------------------------------------------------------------
     disease | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iexposure_1 |    3.006018   .8685449     3.81   0.000     1.706276    5.295827
    constant |     .001001   .0002239   -30.87   0.000     .0006457    .0015519
------------------------------------------------------------------------------
```

Now suppose that we perform a case control study in this population. We recruit data on all 50 cases (S_D =1) and on a sample of 200 controls (S_H=200/29950=0.00667). Ignoring sampling variation, we would expect to obtain (after rounding to whole numbers of controls):

	Exposure		
	+	-	Total
Cases	30	20	50
Controls	67	133	200

The odds ratio estimate = (30 × 133)/(20 × 67) = 2.978, slightly different from the figure above because of rounding.

Fitting a logistic regression model to these case-control data we obtain:

disease	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iexposure_1	2.977612	.9684222	3.35	0.001	1.574084	5.632591
constant	.1503759	.0360648	-7.90	0.000	.0939797	.240615

The estimate of the odds ratio is now 2.978, which is correct apart from rounding errors. The constant parameter has changed substantially, from 0.001001 to 0.1503759. This is because the constant parameter now estimates the true baseline odds (0.001001) multiplied by S_D/S_H = 149.75 (= 0.149899). The small difference is again due to rounding errors.

6. Matched case-control studies

If a case-control study is **broadly matched** (e.g. if we simply ensure that the age distribution is roughly the same in the cases and controls, then logistic regression may be used for analysis **providing the matching variable is included in the model**. Logistic regression cannot be used for the analysis of finely (or individually) matched case-control studies (see future session).

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 10: PRACTICAL

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

(i) use the **logistic/logit** command to estimate the effect (odds ratio) of one exposure, controlled for the effect of a second variable, and obtain a confidence interval for this odds ratio

(ii) use logistic regression to assess if the effect of an exposure (e.g. forest/savannah) is confounded by another variable (e.g. age)

(iii) use the **estimates** and **lrtest** command to perform a statistical test (LRT) of the association between an exposure and the outcome after controlling for the effects of confounding variables.

Using dataset ONCHALL

First we shall investigate whether sex confounds the association between area and mf infection.

1. Produce a cross-tabulation of **area** by **sex**. Is area associated with sex?
2. Tabulate **sex** against **mf** for each area separately. Does sex appear to be associated with mf infection? Do you expect sex to confound the association between area and mf infection? Justify your answer.
3. Perform an analysis of the association between **area** and **mf** stratified on **sex** (using the **mhodds** command). Is it reasonable to calculate a summary odds ratio for the effect of area adjusted for sex? What do you conclude about whether sex is an effect modifier or confounder of the association between area and mf infection?
4. Fit a logistic regression model to obtain an estimate of the odds ratio for area controlled for any confounding effects of sex. How does the odds ratio estimate that you obtain from this analysis compare with that obtained from the Mantel-Haenszel analysis. Does sex appear to be a risk factor for mf infection once area has been taken into account?

Now run the following command:

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

5. What is you interpretation of the odds ratio associated with **_Iarea_1** ? What about the odds ratios associated with **_Isex_1** and **_Iagegrp_2**? Does age confound the association between area and mf infection? Does age confound the association between sex and mf infection?
6. Perform a likelihood ratio test of the null hypothesis that, after controlling for age and sex, there is no association between area and mf infection. What do you conclude? What other statistical test could you have performed of this hypothesis? Does it produce a similar result.

7. Perform a likelihood ratio test of the null hypothesis that, after controlling for area and sex, there is no association between age group and mf infection. What do you conclude? Could you have performed another statistical test of this hypothesis?
8. From preceding analyses, what do you conclude about the relationship between the exposures age, sex and area and the outcome mf infection?

We shall now use logistic regression to analyse a case-control study – the study from Mwanza, Tanzania (`mwanza.dta`). Make sure you have the binary variable `ed2`. If you didn't save it you will need to re-generate it.

9. Tabulate `ed2` against case/control status. What is the crude odds ratio for `ed2`?
10. Make sure Stata knows that `rel`=9 indicates a missing value (`recode rel 9=.`)
11. Using the Mantel Haenszel approach, obtain an estimate of the odds ratio for `ed2` adjusted for religion (`rel`).
12. Use logistic regression to estimate the crude odds ratio for `ed2` and the odds ratio for `ed2` controlling religion. How do your results compare with those you obtained using the Mantel-Haenszel approach.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 11

LOGISTIC REGRESSION 3: INTERACTION

Objectives

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

(i) explain the assumptions which underlie logistic regression models with 2 or more variables in the absence of interaction terms.
(ii) fit logistic regression models that include interaction parameters.
(iii) interpret the parameters that represent interaction/effect modification in regression models.
(iv) perform statistical tests for the presence of interaction.

To control confounding, we have adopted the same approach for cohort studies (week 1), for case control studies (week 2) and in logistic regression modelling (week 3). This approach is to perform the analysis of the effect of exposure of interest *stratifying* according to different levels of the confounding variable(s), so that comparisons are made between individuals who are homogeneous with respect to the confounding variable(s).

In order to combine the information from the different strata, we need to make the assumption that the effect of the exposure is the same across strata (i.e. for different levels of the confounder). Thus in cohort studies, we assumed that the rates for exposed versus unexposed individuals were proportional across strata (i.e. that the true rate ratio is the same in each stratum). For case-control and cross-sectional studies we make the assumption that the odds for exposed and unexposed individuals are proportional across strata (i.e. that the true odds ratio is the same in each stratum).

The `mhodds` and `mhrate` commands give approximate χ^2 tests for interaction/effect modification between the exposure variable and the stratifying variable. These tests are of the null hypothesis the true rate/odds ratios are the same in all strata versus the alternative hypothesis that the true rate/odds ratio (the "effect" of exposure) varies between strata. Any such variation in effect between strata (departure from the proportional rates/odds assumption) is what we call *interaction* or *effect modification*. Thus these tests are of the null hypothesis of no effect modification/interaction versus the alternative hypothesis that there is effect modification/interaction.

In this session we describe how interaction/effect modification can be investigated using regression models.

1. Logistic regression models assuming no effect modification

In the previous session we saw how to estimate the joint effects of two variables (area and age group)in a logistic regression model assuming a constant odds ratio so that we could combine information across strata to produce a summary adjusted estimate of the odds ratio (similar to a Mantel-Haenszel estimate) and perform hypothesis tests. If the odds ratios differ substantially between strata then there is interaction (with respect to odds ratios) between the two variables and the odds ratios for the exposure should be reported separately for different levels of the effect modifying/interacting variable. We shall see how to do this below.

In this session we shall again focus on an analysis of the effects of area and age group on the odds of microfilarial infection. In Exercise 1 of the previous session we calculated the observed odds by area and age group and the stratum–specific odds ratios for area as:

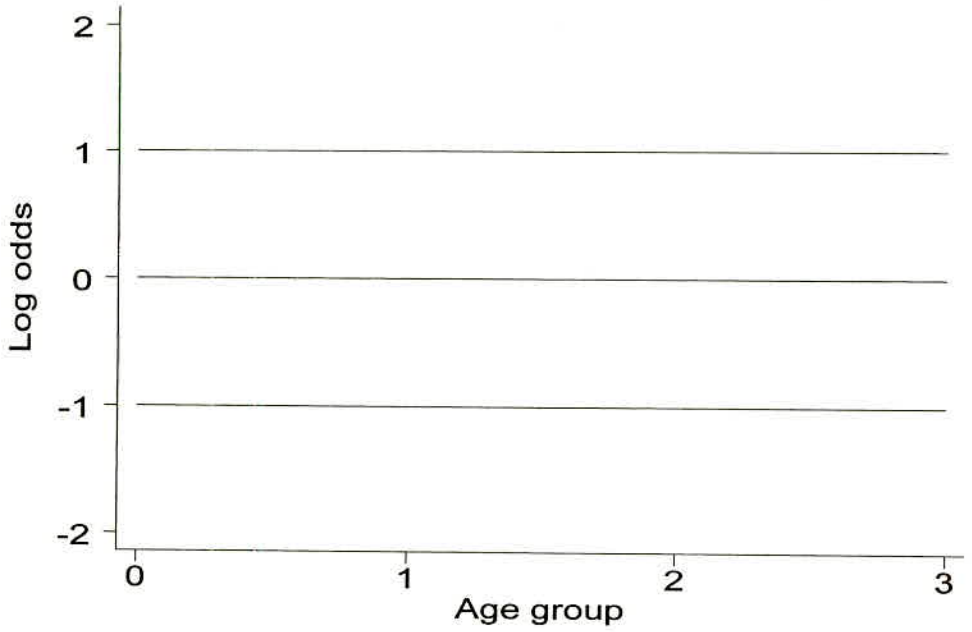
	Area		
Agegrp	0	1	**Odds ratio** (for area)
0	0.208	0.380	1.828
1	0.440	1.116	2.536
2	1.447	4.400	3.041
3	2.182	10.32	4.730

Exercise 1

From the above table of observed odds and odds ratios, calculate the observed log(odds) and log(odds ratios) and complete the table below.

	Area		
Agegrp	0	1	**Log(OR)** (for area)
0	-1.570		
1		0.110	
2	0.369		
3			1.475

Plot the observed log(odds) for each area on the graph below. (Note that we have not made any assumptions about the presence of absence of effect modification in doing these calculations, we are just plotting what we have observed).



In the previous session we fitted the logistic regression model

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

which makes the assumption of no interaction/effect modification, and calculated the fitted odds from the model as:

	Area		
Agegrp	0	1	**Odds ratio**
0	0.147	0.453	3.08
1	0.382	1.178	3.08
2	1.435	4.421	3.08
3	2.593	7.995	3.08

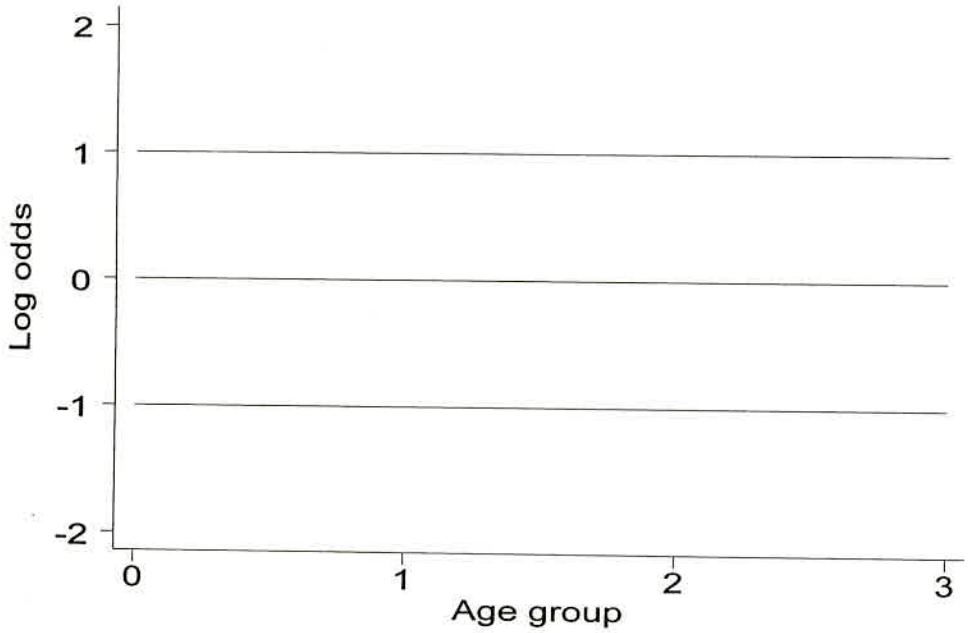
Note that the stratum-specific odds ratios for area are all identical, reflecting the assumption of no interaction between area and age group.

Exercise 2

From the above table of fitted odds and odds ratios, calculate the fitted log(odds) and log(odds ratios) and complete the table below.

	Area		
Agegrp	0	1	**Log(OR)** (for area)
0	-1.917		
1		0.163	
2	0.361		
3			0.125

Plot the fitted log(odds) for each area on the graph below.



Comparing this graph with the previous graph, you should note that the lines areas 0 and 1 are now parallel, which they were not when you plotted the observed log(odds). This is because the distance between the two lines at each point represents the log(odds ratio) at that point. In the model we fitted we assumed constant odds ratios (and hence constant log(odds ratios)) across age groups, thus making the lines are parallel.

When we assume that there is no interaction/effect modification, we assume that:

- The effect of age group is the same in the savannah area as in the forest area, and
- The effect of area is the same in each age group.

To check the assumption of no interaction/effect modification, we can introduce interaction terms into the regression model. To keep things simple initially, we shall recode age into two groups (variable **agebin**): 0= 0-19 years and 1 = 20 or more years

2. Interaction between two binary variables

To investigate whether there is evidence of an interaction, we first tabulate the numbers of infected and uninfected individuals in each age group in each area.

Age group	area variable	**mf**= 1	**mf**= 0	Odds
0-19 (**agebin** = 0)	Savannah (**area** = 0)	38	127	0.299
	Forest (**area** = 1)	107	148	0.723
20 or more (**agebin** = 1)	Savannah (**area** = 0)	243	140	1.736
	Forest (**area** = 1)	434	65	6.677

In order to make it easy to follow some of the steps below, the next table summarises the observed odds for each combination of area and age group (**agebin**).

Age group	**area** = 0	**area** = 1
0-19	0.299	0.732
>=20	1.736	6.677

We can use Stata to calculate a Mantel-Haenszel estimate of the odds ratio for forest versus savannah controlling for age.

```
. mhodds mf area, by(agebin)

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

--------+-------------------------------------------------------------
 agebin | Odds ratio      chi2(1)      P>chi2     [95% Conf. Interval]
--------+-------------------------------------------------------------
   0    | 2.416252        15.84        0.0001     1.542784   3.784244
   1    | 3.846787        67.16        0.0000     2.718532   5.443294
--------+-------------------------------------------------------------

    Mantel-Haenszel estimate controlling for agebin
    ------------------------------------------------------------------
     Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       3.234762      79.68        0.0000         2.462378   4.249423
    ------------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(1)    =    2.60
                                     Pr>chi2    =  0.1072
```

For each age group, we can re-express the odds in **area**=1 in terms of the odds in **area**=0 multiplied by the odds ratio for **area**=1 versus **area**= 0.

Age group	**area** = 0	**area** = 1	
0-19	0.299	0.732	= **0.299 × 2.416**
>=20	1.736	6.677	= **1.736 × 3.847**

Note that the observed odds ratios for the effect of area are somewhat different in the two age groups (2.4 versus 3.8). The assumption of no interaction states that the *true* odds ratios in each age group are equal. Under this assumption, the estimated odds ratio for forest versus savannah (**area**=1 vs **area**=0), controlling for the confounding effects of age, is 3.23. This is a weighted average of the stratum-specific odds ratios (2.416 and 3.847).

Next, we present the table with the odds in those aged 20+ further expanded in terms of the odds in those aged 0-19 years (0.299) and the odds ratio for age 20+ versus age group 0-19 (1.736/0.299 = 5.801):

Age group	**area** = 0			**area** = 1			
0-19	0.299			0.723	=	0.299	× 2.416
>=20	1.736	=	**0.299 × 5.801**	6.677	=	**0.299 × 5.801**	× 3.847

Finally, we can rewrite the odds ratio for area in the older age group (3.847) in terms of the odds ratio for area in the baseline age group multiplied by an interaction parameter (3.847 = 2.416 × 1.592).

Age group	**area** = 0 (savannah)			**area** = 1 (forest)			
0-19	0.299			0.723	=	0.299	× 2.416
>=20	1.736	=	0.299 × 5.801	6.667	=	0.299 × 5.801	× **2.416 × 1.592**

Note that the odds in the four combinations of area and age group are now expressed in terms of 4 parameters:

0.299	=	odds in the baseline group (**area**=0 and **agebin**=0)
2.416	=	odds ratio for area in the baseline age group (0-19)
5.801	=	odds ratio for age group in the baseline area (savannah)
1.592	=	the interaction parameter – which quantifies the extent to which the odds ratios for area and age group in the non-baseline groups differ from those in the baseline groups.

I.e. odds ratio for area 1 versus area 0 =		2.416 in the baseline age group (0-19)
	=	2.416 × 1.592 in the non-baseline age group.
OR for age>=20 versus age=0-19	=	5.801 in the baseline area (savannah)
	=	5.801 × 1.592 in the non-baseline area (forest)

To fit the model allowing interaction/effect modification in Stata, we simply place an asterisk (*) between the variables in the logistic or logit commands.

```
. xi: logit mf i.area*i.agebin
i.area            _Iarea_0-1          (naturally coded; _Iarea_0 omitted)
i.agebin          _Iagebin_0-1        (naturally coded; _Iagebin_0 omitted)
i.area*i.agebin   _IareXage_#_#       (coded as above)

Iteration 0:   log likelihood = -902.47763
Iteration 1:   log likelihood =  -712.7589
Iteration 2:   log likelihood = -707.07326
Iteration 3:   log likelihood = -706.99057
Iteration 4:   log likelihood = -706.99054

Logit estimates                                   Number of obs   =       1302
                                                  LR chi2(4)      =          .
Log likelihood = -706.99054                       Prob > chi2     =          .
```

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.416252   .5418676     3.93   0.000     1.556869    3.750009
  _Iagebin_1 |    5.80094   1.236674     8.25   0.000     3.819749    8.809715
_IareXage_~1 |   1.592047   .4481527     1.65   0.099     .9169535    2.764169
    constant |   .2992126   .0553259    -6.53   0.000     .2082517    .4299037
------------------------------------------------------------------------------
```

The point estimates obtained from this model correspond to the numbers we calculated in the table above. This model does not assume that the effect of area is the same in both age groups (or that the effect of age is the same in both areas).

Interpretation

The interpretation of the parameters in this model is as follows:

- Among people aged 0 – 19 years (i.e. at the baseline of `agebin`) the odds ratio for the effect of area (forest versus savannah) is 2.42, (**`_Iarea_1`**).
- Among people living in the savannah (i.e. at the baseline of `area`) the odds ratio for the effect of age (20+ versus 0-19) is 5.80, (**`_Iagebin_1`**).
- Among people aged 20+ years (i.e. not at the baseline of `agebin`) the odds ratio for the effect of area (forest versus savannah) is 2.42 multiplied by the interaction parameter 1.59 (**`_IareXage_~1`**).
- Among people living in the forest (i.e. not at the baseline of `area`) the odds ratio for the effect of age (20+ versus 0-19) is 5.80 multiplied by the interaction parameter 1.59 (**`_IareXage_~1`**).
- The odds (not odds ratio) of infection in those in the baseline of both area (savannah) and age group (0-19) is 0.299 (**`constant`**)

Question: What is the true value of the interaction parameter if there is no interaction?

Very Important Note: In the current model, with an interaction term between **`area`** and **`agebin`**, the parameter **`_Iarea_1`** is interpreted as a stratum-specific odds ratio, the odds ratio for area in one stratum of age. In the model fitted in the previous session assuming no interaction (**`xi:logit mf i.area i.agegrp`**), the parameter **`_Iarea_1`** was interpreted as the summary odds ratio for area adjusted for the effect of age.

3. Variables with more than two levels

Returning to age classified into 4 groups, we can write down the odds of infection in each age group/area combination as follows (see previous session).

	Area	
Age group	0	1
0	0.208	0.380
1	0.440	1.116
2	1.447	4.400
3	2.182	10.32

If we express this table in terms of the odds ratios for age groups 1 to 3 compared to age group 0, and the odds ratios for area 1 versus area 0, we get:

Age group	Area 0	Area 1
0	0.208	0.208 × 1.828
1	0.208 × 2.118	0.208 × 2.118 × 2.536
2	0.208 × 6.964	0.208 × 6.964 × 3.041
3	0.208 × 10.50	0.208 × 10.50 × 4.730

Assuming no interaction, we calculated that the odds ratio for Area 1 (forest) versus Area 0 (savannah), controlling for the confounding effects of age, was 3.08.

By introducing interaction parameters (in bold), we express the odds in each group in terms of the odds ratio for each variable at the baseline of the other and the interaction parameters:

Age group	Area 0	Area 1
0	0.208	0.208 × 1.828
1	0.208 × 2.118	0.208 × 2.118 × 1.828 × **1.388**
2	0.208 × 6.964	0.208 × 6.964 × 1.828 × **1.664**
3	0.208 × 10.50	0.208 × 10.50 × 1.828 × **2.588**

The odds in the eight combinations of area and age group are now expressed in terms of 8 parameters:

0.208 = odds in the baseline group (**area**=0 and **agegrp**=0)
1.828 = odds ratio for area in the baseline age group (0-9)
2.118 = odds ratio for age group 1 vs age group 0 in baseline area (savannah)
6.964 = odds ratio for age group 2 vs age group 0 in baseline area (savannah)
10.50 = odds ratio for age group 3 vs age group 0 in baseline area (savannah)
1.388 = the interaction parameter which quantifies the extent to which the odds ratio for area in age group 1 differs from the odds ratio for area in age group 0 and the extent to which the odds ratio for age group 1 versus age group 0 differs in the forest from that in the savannah.
1.664 = the interaction parameter which quantifies the extent to which the odds ratio for area in age group 2 differs from the odds ratio for area in age group 0 and the extent to which the odds ratio for age group 2 versus age group 0 differs in the forest from that in the savannah.
2.588 = the interaction parameter which quantifies the extent to which the odds ratio for area in age group 3 differs from the odds ratio for area in age group 0 and the extent to which the odds ratio for age group 3 versus age group 0 differs in the forest from that in the savannah.

Exercise 3

What is the odds ratio for age group 2 vs age group 0 in the forest area (**area**=1)?
What is the odds ratio for area in age group 3?

Fitting this model in Stata:

```
. xi: logistic mf i.area*i.agegrp
Logit estimates                                   Number of obs   =       1302
                                                  LR chi2(8)      =          .
Log likelihood = -689.77289                       Prob > chi2     =          .

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   1.827532   .6369638     1.73   0.084     .9229731    3.618602
  _Iagegrp_1 |     2.1175   .7949557     2.00   0.046     1.014527    4.419601
  _Iagegrp_2 |   6.963971   2.150749     6.28   0.000     3.801652    12.75679
  _Iagegrp_3 |       10.5   3.353467     7.36   0.000     5.614802     19.6356
_IareXage_~1 |   1.387791    .642613     0.71   0.479     .5599863    3.439305
_IareXage_~2 |   1.663802   .6901372     1.23   0.220     .7379504     3.75125
_IareXage_~3 |    2.58819     1.1337     2.17   0.030     1.096847    6.107261
    constant |   .2077922   .0570907    -5.72   0.000     .1212725    .3560377
------------------------------------------------------------------------------
```

The parameter estimates correspond exactly with the terms in the table above. On a log scale we obtain:

```
. logit
```

```
------------------------------------------------------------------------------
          mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iarea_1 |   .6029662   .3485378     1.73   0.084    -.0801552    1.286088
  _Iagegrp_1 |   .7502361   .3754218     2.00   0.046     .0144229    1.486049
  _Iagegrp_2 |    1.94075   .3088394     6.28   0.000     1.335436    2.546064
  _Iagegrp_3 |   2.351375   .3193778     7.36   0.000     1.725406    2.977344
_IareXage_~1 |   .3277132   .4630474     0.71   0.479     -.579843    1.235269
_IareXage_~2 |   .5091052   .4147953     1.23   0.220    -.3038787    1.322089
_IareXage_~3 |    .950959   .4380289     2.17   0.030     .0924381     1.80948
       _cons |  -1.571217    .274749    -5.72   0.000    -2.109715   -1.032719
------------------------------------------------------------------------------
```

From this output, we can obtain the log(odds) fitted by the model from the equation:

log(odds) = -1.571
+ 0.603×_Iarea_1
+ 0.750×_Iagegrp_1 + 1.941×_Iagegrp_2 + 2.351×_Iagegrp_3
+ 0.328×_IareXage_~1 + 0.509×_IareXage_~2 + 0.951×_Iareage_~3

Exercise 4

What are the fitted log odds for the following individuals:

- someone in the savannah (**area**=0) in age group 2
- someone in the forest (**area**=1) in age group 0
- someone in the forest (**area**=1) in age group 3

How do the fitted log odds compare with the observed log odds that you calculated in Exercise 2?

4. Likelihood ratio test for interaction

We can perform a likelihood ratio test of the null hypothesis that there is no interaction between area and age group versus the alternative hypothesis that there is an interaction. This test is analogous to the "Test of homogeneity of ORs" performed by the **mhodds** command. We do this by fitting models with and without interactions terms and comparing the log likelihoods for the two models.

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

Log likelihood = -689.77289

estimates store A

xi: logistic mf i.area i.agegrp

Log Likelihood = -692.40733
```

```
estimates store B

lrtest B A

likelihood-ratio test                                 LR chi2(3)  =       5.27
(Assumption: B nested in A)                           Prob > chi2 =     0.1531
```

Note that: $5.27 = 2 \times (-689.773 + 692.407)$,
the test is on 3 degrees of freedom because there are 3 interaction parameters being tested simultaneously,
the test statistic (5.27) is similar to that provided with the **mhodds** command (5.05, also 3 degrees of freedom, $p = 0.17$).

The result of the likelihood ratio test indicates that the data are compatible with the assumption of no interaction/effect modification (p=0.15). Note that if we look at the individual Wald tests for the three interaction parameters, there is one which has a p-value of 0.03, while the others have p-values of 0.22 and 0.48. Assessing the true significance of the individual Wald tests is complicated because we are performing multiple tests. Therefore it is better to concentrate on the results of the likelihood ratio test.

5. Interpreting interaction terms

In the above example, the estimated interaction parameters are all positive (on the log odds scale) or greater than 1 (on the odds scale). If we ignore the results of the LRT above, and assume that interaction is present, then the fact that the baseline odds ratios are greater than 1 and the interaction parameters are also greater than 1 and increasing with age, implies that the effect of area increases with age (and vice versa). If we plot the observed log odds (Exercise 2), we see that the lines diverge as age increases.

If the interaction terms had been negative, this would have corresponded to lines which converged with increasing age.

If there is interaction, it is no longer appropriate to report the effect of an exposure adjusted for the confounder, since the proportional odds assumption which is used to combine information across strata is incorrect. Instead, we should report the **stratum-specific** exposure effects. A way to obtain stratum-specific odds ratios with confidence intervals is introduced in the practical to this session. The results are as follows:

Age group	**Odds ratio (95% CI) for savannah v forest**
0 (0-9)	1.83 (0.92, 3.62)
1 (10-19)	2.54 (1.40, 4.61)
2 (20-39)	3.04 (1.96, 4.72)
3 (40+)	4.73 (2.81, 7.96)

An advantage of classical (Mantel-Haenszel) methods is that we are reminded and encouraged to look for evidence of interaction. In regression models there is no reminder and we have to fit interaction terms explicitly to do this.

6. More complicated models

All the examples given in lectures up to this point have been of logistic regression models with one or two explanatory variables. These have enabled us to see exactly how the parameter estimates are derived, and to explain the link with classical methods of analysis. The power of regression models, however, is that they allow us to examine simultaneously the effects of a number of variables, in situations where classical methods become impractical. The 'cost' of this 'power' is in the extra assumptions made about how the effects of the variables combine.

Example

Suppose we wish to examine the effect of area of residence on microfilarial infection, adjusting for the potential confounding effects of both age group and sex. Using classical methods, we stratify the analysis by both the confounders giving eight strata (4 for age group × 2 for sex). Using STATA:

```
mhodds mf area, by (agegrp sex)
```

```
------------------------------------------------------------------------------
   agegrp          sex | Odds Ratio     chi2(1)        P>chi2     [95% Conf.
Interval]
--------------------+---------------------------------------------------------
        0            0 |   1.388430        0.49       0.4861      0.54916
    3.51033
        0            1 |   2.852174        3.73       0.0536      0.93669
    8.68473
        1            0 |   1.769585        1.74       0.1866      0.74996
    4.17546
        1            1 |   3.243421        6.70       0.0096      1.26296
    8.32946
        2            0 |   2.619048        6.03       0.0141      1.17850
    5.82048
        2            1 |   3.223881       18.42       0.0000      1.83110
    5.67605
        3            0 |   2.627451        7.82       0.0052      1.29983
    5.31106
        3            1 |   9.725490       35.36       0.0000      3.86748
   24.45654
------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for agegrp and sex
    ------------------------------------------------------------
     Odds Ratio     chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------
       3.011474       67.30        0.0000         2.283029   3.972344
    ------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(7)    =    11.63
                                     Pr>chi2    =   0.1136
```

In general, the number of strata is the product of the number of strata for each of the confounders. This leads to large numbers of strata, with (usually) small amounts of data in each stratum, when there are more than two or three confounders.

We can perform a logistic regression analysis equivalent to the Mantel-Haenszel analysis above by fitting the model:

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

and performing a likelihood ratio test by omitting **area** in the usual way. The result for the effect of area only (we have omitted the Stata output for the other variables) is:

```
 mf   | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------+------------------------------------------------------------------
area  | 3.059568    .4254645     8.042   0.000     2.329656    4.018171
(... other parameters)

Logistic:  likelihood-ratio test                  chi2(1)      =      68.27
                                                  Prob > chi2  =     0.0000
```

The Mantel-Haenszel results are close to the maximum likelihood estimates given by logistic regression.

However, the logistic regression model can be simplified, providing that there is no interaction between the two confounding variables. If we fit the model assuming no interaction between any of the three variables:

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

then the results are very similar:

```
  mf  | Odds Ratio Std. Err.      z    P>|z|     [95% Conf. Interval]
------+------------------------------------------------------------------
 area | 3.073138    .425981     8.099   0.000     2.342034    4.032467
(... other parameters)

Logistic:  likelihood-ratio test                  chi2(1)      =      69.36
                                                  Prob > chi2  =     0.0000
```

The estimated effect of **area** is almost identical. However this model contains three less parameters, since the age/sex interaction terms are omitted. Therefore the regression approach is more 'efficient' than the Mantel-Haenzsel approach as it involves fewer parameters. This increased efficiency is gained at the cost of extra assumptions (that the effects of age and sex combine multiplicatively).

Suppose we had four confounders, with 2, 3, 4 and 5 levels respectively. The number of strata, and of parameters in a logistic regression model which included all interactions between them, would be $2\times3\times4\times5 = 120$. However a logistic regression model which assumes no interaction contains only $1+1+2+3+4 = 11$ parameters. For more complicated models, the simplifying assumption of no interaction between the effects of the different variables allows us to examine joint (simultaneous) effects which would be impossible to estimate using classical methods.

As a general rule, interactions between confounders, especially those with many levels are omitted from regression models.

7. Increasing power in tests for interaction

Suppose that we wish to test for interaction between two variables, one of which has r levels and the other c levels. The number of interaction parameters will be $(r - 1)\times(c - 1)$. This means that tests for interaction can involve a large number of parameters and, as a result, have **low power**. That is, they may not provide evidence of interaction even when it is present.

Another problem is that some strata may have no individuals or no cases. If this is the case, then interaction parameters for that stratum cannot be estimated. Stata deals with this by dropping such strata and parameters, but this means that the model including the interaction term is not directly comparable with the one assuming no interaction.

A solution to both these problems is to **combine groups**, so that the interaction introduces only a small number of extra parameters. Thus:

a) testing for interaction between area of residence and age group as originally coded (4 groups) gives $\chi^2= 5.27$ on 3 degrees of freedom, p=0.1531

b) testing for interaction between area of residence and age group coded as 0-19 and 20 gives $\chi^2= 2.70$ on 1 degree of freedom, p=0.10

A fuller account of the use of regression models to examine interaction is given in Clayton & Hills, pp239-242 and Chapter 26.

Appendix Example of a Do-file to produce graphs of log(odds)

```
*     grlodds.do

*     programme to graph log odds for oncho data

version 8.0

use onchall,clear
preserve

* choose model you want to plot

quietly {

  xi:logit mf i.area*i.agegrp
  predict mfpred

  collapse (sum) pos=mfpred (count) denom=mf,by(area agegrp)
  gen logodds=log(pos/(denom-pos))

  gen lodds0=logodds if area==0
  gen lodds1=logodds if area==1

  label variable lodds0 "Log odds, savannah"
  label variable lodds1 "Log odds, forest"

}

twoway (scatter lodds0 agegrp, c(l) s(T)) /*
     */(scatter lodds1 agegrp, c(l) s(+)),xlab(0 1 2 3)

restore
```

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 11 PRACTICAL

Logistic regression 3

By the end of this practical students will be able to:

i) investigate interaction between two variables using logistic regression in STATA
ii) interpret the interaction parameters produced by STATA
iii) conduct a likelihood ratio test for interaction
iv) understand that the power of a test for interaction may be increased by combining groups if one variable has several categories

1. Using dataset ONCHALL create the binary age variable **agebin**:

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

Check the new variable by cross-tabulating it against the old one.

Use STATA to fit a logistic regression model of the effects of area, binary age group and their interaction, on the odds of microfilarial infection.

Write down the following odds ratios from the output:

i) Odds ratio for area (forest versus savannah) at the baseline value of age.

ii) Odds ratio for age (20+ versus 0-19) at the baseline value of area (savannah)

iii) the interaction term (odds ratio scale) between age and area

What is the mf infection odds ratio for individuals aged 20+ living in the forest (area=1) versus individuals aged 0-19 living in the forest? I.e. what is the effect of age in the forest?

2. Now fit a logistic regression model for the effects of area with age group in four categories and their interaction, on the odds of microfilarial infection.

Write down the following odds ratios from the output:

i) The odds ratio for the effect of area at the baseline value of age.

ii) The odds ratios for the effect of age group at the baseline value of area (savannah).

`agegrp` = 1	`agegrp` = 2	`agegrp` = 3

iii) the interaction terms between age and type of area.

area(1).agegrp(1)	area(1).agegrp(2)	area(1).agegrp(3)

By combining the appropriate odds ratios, derive the 4 age-specific odds ratios for area [response table below]

Age category	OR (forest vs savannah)
Agegroup 0	
Agegroup 1	
Agegroup 2	
Agegroup 3	

vi) use the `lincom` command to obtain directly from STATA the age-specific odds ratios for the effect of area together with a 95% confidence interval. Check the STATA output against the odds ratios you calculated in part vi). (Note: Be careful to check the names Stata has given to the dummy interaction variables/ parameters)

e.g. `lincom _Iarea_1 + _IareXage_1_1,or`

3. Use STATA to test for interaction between type of area and age, first using the four-category variable `agegrp` and then using the binary variable `agebin`. In each case first run a model including the interaction term and then run one without and perform an LR test. What do you notice about the p-values? What is the interpretation of the parameter `_Iarea_1` in the models with interaction terms between age and area? What is its interpretation in the models without any interaction term?

	LR test for interaction: $\chi 2$; df	p-value
agebin		
agegrp		

4. A problem with combining age categories is that our model now fails to allow adequately for the change in the odds of disease with age which, as we saw in previous sessions, varies significantly between different levels of **agegrp**.

   A way round this is to include the original variable **agegrp** as a stand alone "main effect" in the model and to include also area, agebin and the interaction between them.

   Fit this model:

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

   What is the estimated effect (odds ratio) of area in those aged 0-19? What is the estimated effect of area in those aged 20+?

   To test the statistical significance of the interaction save the log likelihood for the model above, fit the model

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

   and perform a likelihood ratio test.

   Important note: it can be tempting to try different combinations of categories or to combine categories after looking at stratum-specific odds ratios to see whether there appear to be differences. The problem with this approach is that it will increase the frequency with which "false positive" interactions are identified. In a situation in which there is no interaction, trying different cut-offs or choosing a cut-off after inspection of the data will increase the chances of a statistically significant result even though no interaction is present.

5. In this session we are focussing on area as exposure and microfilarial infection as outcome and trying to get the best measure of the effect of area on infection. There is little point in having a complex model which is difficult to interpret if a simpler one is statistically almost as good.

   Let us compare two models which include age group and sex as possible confounders for the effect of type of area on microfilarial infection. In one we simply include the two factors, **agegrp** and **sex**, as main effects. In the other we also add in an interaction term between **agegrp** and **sex**. We want to see whether the more complex model leads to a different estimate of the effect of type of area than the simpler model and whether it alters our views of **area** being a risk factor for microfilarial infection.

Fit the model

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

Note the odds ratio for the effect of type of area on the outcome. Is there evidence for an association between area and odds of infection after controlling for age and sex?

Repeat the exercise but this time with the model

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

Are the estimates of the area parameter different in the two models and has the p-value for the significance of area changed?

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 12

Logistic regression with quantitative measures of exposure

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

(i) fit and interpret logistic regression models with quantitative measures of exposure.
(ii) explain the difference in fitting ordered categorical exposure variables as either a quantitative or a categorical variable, and be able to choose the most appropriate approach given the circumstances.
(iii) describe, qualitatively, key similarities and differences between logistic and other regression methods.

So far, we have only used logistic regression to examine the effects of categorical exposure variables. To do this, we tell STATA to create indicator variables corresponding to the different exposure levels.

One of the ways in which regression models allow us to examine exposure effects more flexibly than classical methods is that the effect of variables which are either continuous or ordered categorical can be examined. For such variables, it is possible that any exposure effect will increase (or decrease) systematically with the level of exposure. This is known as a dose-response relationship, or trend. The demonstration of such a relationship provides more convincing evidence of a causal effect of exposure than a simple comparison of exposed with unexposed subjects (one of Bradford Hill's criteria).

1. Review of simple linear regression

Those of you who studied statistics during Term 1 probably encountered the use of linear regression to examine the association between a continuous response variable (y) and an exposure variable (x). Linear regression assumes a relationship of the form

$$y = a + bx$$

and estimates the parameters a (called the intercept) and b (the slope or gradient). The intercept is the estimated value of y when x=0, while the slope is the estimated increase in y per unit increase in x.

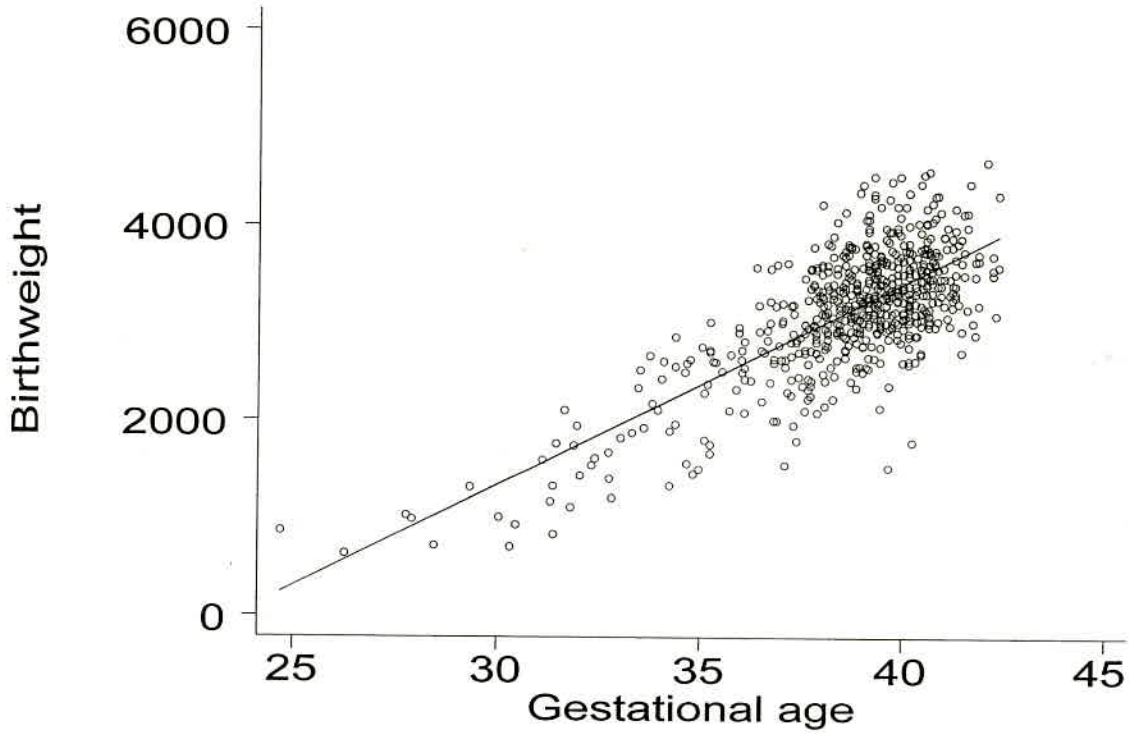
For example, we can use linear regression to examine the association between gestational age and birthweight. (Before actually fitting a regression model to such data we should produce a scatter plot to see whether it is sensible to fit a linear regression model – see scatter plot below.) Fitting a linear regression model produces the following STATA output:

bweight	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gestwks	206.6412	7.484572	27.609	0.000	191.9439	221.3386
_cons	-4865.245	290.0814	-16.772	0.000	-5434.873	-4295.617

This tells us that the estimated regression equation is:

$$\texttt{bweight} = -4865 + 206.6 \times \texttt{gestwks}$$

We interpret the slope parameter, 206.6, as the estimated increase in birthweight for each additional one week of gestation.



2. Logistic regression with quantitative exposure variables

A logistic regression model with a quantitative exposure variable can be written in the form:

$$\log(\text{odds}) = a + bx$$

Here a (the intercept) corresponds to the corner parameter in our previous model and represents the log(odds) in individuals whose value of x is 0. b is the regression coefficient for variable x and represents the common change in log(odds) per unit change in x. In Kirkwood and Sterne (page 203) this model is written as:

$$\log(\text{odds}) = \text{Baseline} + [X]$$

where [] indicate the assumption that the exposure X has a linear effect on the log(odds).

Example: the effect of age on the odds of microfilarial infection.

We return (for the last time!) to the association between age and microfilarial infection.

	mf			
agegrp	0	1	odds	log(odds)
0	156	46	0.29	-1.221
1	119	99	0.83	-0.184
2	125	299	2.39	0.872
3	80	378	4.73	1.553

Fitting age group as a categorical variable (as we have done in past sessions) we obtain the following results:

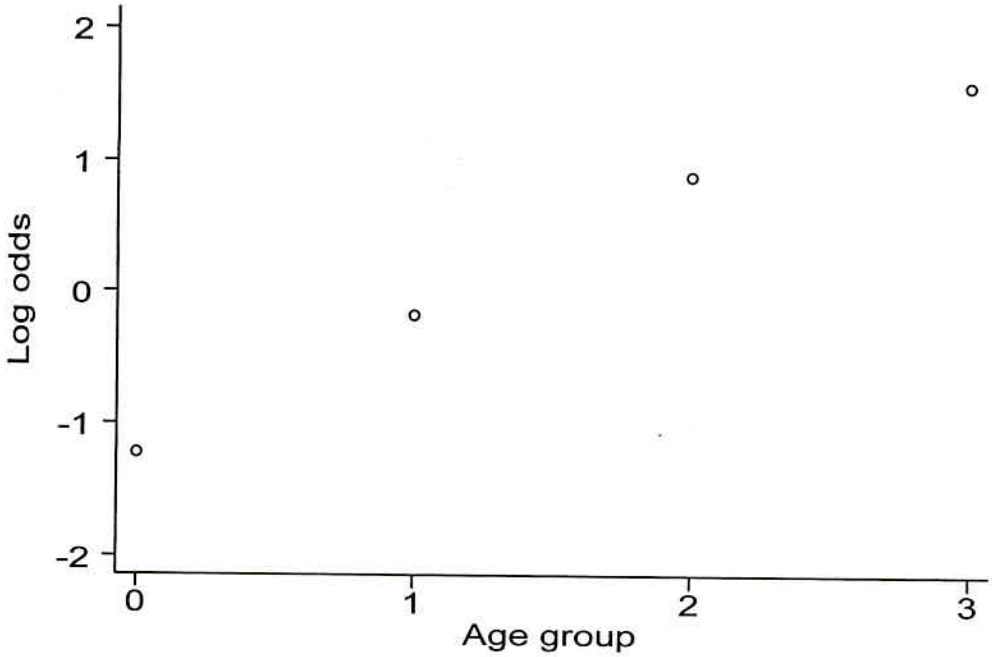
```
xi: logistic mf i.agegrp
i.agegrp          _Iagegrp_0-3        (naturally coded; _Iagegrp_0 omitted)

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

------------------------------------------------------------------------------
          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 |      8.112   1.612103    10.53   0.000     5.495005    11.97534
  _Iagegrp_3 |   16.02391   3.334152    13.33   0.000     10.65752    24.09246
------------------------------------------------------------------------------
```

or, on the log(odds) scale:

```
------------------------------------------------------------------------------
          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   -.8923762
------------------------------------------------------------------------------
```

Exercise 1

The graph below plots the log(odds) of infection for each age group. From the Stata output the left-most point is just the **_cons** coefficient. The next point is **_cons+_Iagegrp_1**, etc. Draw a straight line 'by eye' through these points and estimate the slope of the line you have drawn.



We now examine the estimated effect of age treating the value of **agegrp** as quantitative. The simplest such model is:

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

This model assumes that the increase in log(odds) is the same for each unit increase in the value of **agegrp**. I.e. the log(odds) increase (or decrease) by the same amount when we move from age group 2 to age group 3 as when we move from age group 0 to age group 1. The regression coefficient for **agegrp** gives the maximum likelihood estimate of this common change in log(odds) per unit change in age group (the linear effect of age group on the log(odds)).

```
Logit estimates                                   Number of obs   =       1302
                                                  LR chi2(1)      =     255.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -729.23967                       Pseudo R2       =     0.1491

------------------------------------------------------------------------------
          mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   .9297561   .0637378    14.59   0.000     .8048322     1.05468
       _cons |  -1.114923   .1269573    -8.78   0.000    -1.363754    -.866091
------------------------------------------------------------------------------
```

This output indicates that, the maximum likelihood estimate of the slope of the line that you tried to draw by eye is 0.93. How close was your estimate?

Note: STATA assumes that you wish to treat explanatory variables as quantitative unless you tell it to use indicator variables. Fitting such a model is only appropriate if the values the variable takes represent something quantitative. The values of the variable should ideally reflect the degree of exposure. For example, if the exposure was level of blood pressure and

the four categories of exposure were obtained by grouping the blood pressures, then the values might be the midpoints or the mean of blood pressure in each of the four groups.

The model we have fitted can be written as:

$$\log(\text{odds}) = \text{Baseline} + [\text{Agegrp}]$$

and the estimated regression line obtained from this model as:

$$\log(\text{odds}) = -1.1149 + (0.9298 \times \texttt{agegrp})$$

The fitted log(odds) in each age group are then:

Age group	**log(odds)**
0	-1.1149
1	-1.1149 + 0.9298×1
2	-1.1149 + 0.9298×2
3	-1.1149 + 0.9298×3

Notes:

- The log(odds) increase by 0.9298 for each unit increase in the variable **agegrp**. So, 0.9298 is the estimate of the log(odds ratio) per unit increase in **agegrp**.
- The model **assumes** that the log(odds) increase in a linear manner with **agegrp**.
- With this assumption, focusing our attention on the odds rather than the log(odds), we see that the odds are **multiplied** by exp (0.9298) = 2.53 each time we move from one exposure level (age group) to the next.

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   2.533891   .1615047    14.59   0.000     2.236321    2.871056
------------------------------------------------------------------------------
```

Exercise 2

Calculate the fitted log(odds) for this model, for each value of agegrp.

Age group	**fitted log(odds)**
0	
1	
2	
3	

Plot these fitted log(odds) (i.e. the regression line) on the graph for Exercise 1. What is the log(odds ratio) associated with a change of 2 age groups?

3. Testing for departure from a linear trend.

In the output above the results of a Wald test are presented ($z=14.59$, $p<0.001$). The null hypothesis being tested is that there is no association between age group and odds of infection; the alternative hypothesis is that there is a linear association between age group and the log(odds) of infection. The alternative hypothesis itself is quite restrictive. Why should the association be linear?

We have now fitted two models for the effect of age group on microfilarial infection; firstly with age group as a categorical variable, allowing the possibility of a non-linear relationship between age group and the log(odds) of infection, and secondly with age group as a quantitative variable, assuming a linear relationship. In this instance, the fitted log(odds) for the model assuming a linear trend with age group are close to the observed log(odds). Because of random variation, however, the fitted log(odds) will never exactly match the observed log(odds). In general, we will wish to know whether the assumption of a linear increase in log(odds) with exposure is reasonable.

To do this we will test the null hypothesis that the log(odds) of mf infection increase linearly with the value of the variable **agegrp**. The alternative hypothesis is that there is extra variation in the log(odds) of disease beyond that accounted for by the linear trend. Put another way, the null hypothesis is that the relationship between age group and log(odds) of infection is linear, the alternative hypothesis is that the relationship is more complicated than that.

Up until now, we have performed likelihood ratio tests by fitting two models, one with and one without the variable of interest. In this instance, we shall compare:

1. Baseline + Agegrp (`xi:logistic mf i.agegrp`)

with

2. Baseline + [Agegrp] (`logistic mf agegrp`)

In order for the likelihood ratio test to be valid, the second (simpler) model **must** be a special case of the first (a "nested" model). This is obviously true when we restrict a model by omitting a variable - the second is a special case of the first where the parameter(s) (log(odd ratios)) for the variable of interest = 0.

This is also the case here - the second model is a special case of the first in which:

_Iagegrp_2 = 2×**_Iagegrp_1**
_Iagegrp_3 = 3×**_Iagegrp_1**.

```
xi: logistic mf i.agegrp

Log Likelihood = -727.83149

estimates store A

logistic mf agegrp

Log Likelihood = -729.23967

estimates store B
lrtest B A

likelihood-ratio test                           LR chi2(2)  =      2.82
(Assumption: B nested in A)                     Prob > chi2 =    0.2446
```

The difference in the degrees of freedom is the difference between the number of parameters with age group as a quantitative variable (1+1) and as a categorical variable (1+3). The likelihood ratio test result indicates that the data are compatible with the null hypothesis that the log(odds) of infection increase linearly with the value of `agegrp`.

4. Models for the effect of more than one variable

So far, we have looked at the effect of a single exposure. The considerations are exactly the same for models including two or more variables: the interpretation now is that estimated odds ratios are controlled for the effects of the other variables in the model, on the assumption of no interaction between them (unless interaction terms are included in the model).

The general form of a logistic regression model for the effect of exposure variables x_1, x_2, x_3 etc. is:

$$\log(\text{odds}) = a + b_1x_1 + b_2x_2 + b_3x_3 + ...$$

- If x_i is an binary (indicator) variable (1=exposed, 0=unexposed), then b_i (the coefficient for x_i) is the log(odds ratio) for exposed compared to unexposed individuals, controlling for the effects of the other variables.

- More generally, b_i is the increase in log(odds) per unit increase in x_i, controlling for the effects of the other variables.

5. Variables measured on a continuous scale

Up to now, we have used age divided into 4 categories (5-9, 10-19, 20-39, 40+). Age is often measured more accurately than this. It is of course possible to use logistic regression to model the effect of age in measured in years if such data are available.

With a binary response variable it is not possible to plot the relationship between the odds of disease and the continuous exposure directly, since the odds for each individual are either 0 or infinity. It is thus very important to check for departure from linearity when using continuous variables in logistic regression models. This can be done by grouping the variable and examining the odds of disease in each group, and by introducing quadratic (or higher power) terms (see below).

With a variable such as age, which is often strongly associated with disease and must be considered as a potential confounder in most studies, the standard practice in epidemiological studies, rather than using age in years, is to create 5 or 10 year age groups, which are then used as a categorical (indicator) variable in modelling. The assumption is that the odds of disease do not vary greatly within each group. Although this introduces more terms into the model, it avoids making any assumptions about the precise relationship between age and the odds of disease.

6. Final remarks

These four sessions on logistic regression serve as an introduction not just to regression modelling, but to a range of regression models which are commonly used in epidemiological research. For instance, we use:

- Conditional logistic regression to analyse finely matched case-control studies.
- Poisson regression to analyse cohort studies.
- Cox (Proportional Hazards) regression to analyse cohort studies with rapidly varying rates.

The principles of modelling that you have met in this introduction to logistic regression apply to these other regression models as well. If you feel comfortable fitting logistic regression models then you should find it very easy to:

- interpret parameter estimates
- control confounding
- test for interaction
- fit linear trends
- perform likelihood ratio tests

using conditional logistic, Poisson or Cox regression models.

For more on regression modelling, see Kirkwood and Sterne (various chapters) or Part II of Clayton & Hills or attend the study unit Advanced Statistical Methods in Epidemiology (ASME).

The following material is supplementary and will not be covered in the lecture.

7. Quadratic dose-response relationships.

The simplest departure from a linear relationship between the log(odds) and a quantitative exposure is a quadratic relationship. To examine this, we create a new variable whose values are the squares of the original exposure values.

```
gen agegrp2=agegrp*agegrp
```

agegrp	agegrp2
0	0
1	1
2	4
3	9

Another way to test for departure from a linear trend in the log(odds), particularly when there is a large number of categories, is to fit the model

$$\log(\text{odds}) = \text{Baseline} + [\text{Agegrp}] + [\text{Agegrp2}]$$

i.e.

$$\log(\text{odds}) = a + b_1 \times \texttt{agegrp} + b_2 \times \texttt{agegrp}^2$$

by typing **`logistic mf agegrp agegrp2`**

and then omit **agegrp2** and perform a likelihood ratio test. This provides a test on 1 degree of freedom which will be sensitive to a departure from linearity in which the effects of age increase or decrease with age. **Note:** you must include both the linear and quadratic terms in the quadratic model.

Example — test for extra-linear effect of age:

1. Fit the model including **agegrp2**

`logistic mf agegrp agegrp2`

```
Log Likelihood = -728.08125

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   3.468107   .7563043     5.70   0.000     2.261869    5.317622
     agegrp2 |   .9046999   .0596808    -1.52   0.129     .7949739    1.029571
------------------------------------------------------------------------------
```

`logit`

```
------------------------------------------------------------------------------
          mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   1.243609   .2180741     5.70   0.000      .8161913    1.671026
     agegrp2 |   -.100152   .0659675    -1.52   0.129     -.2294459     .029142
       _cons |  -1.257264   .1614159    -7.79   0.000     -1.573633   -.9408948
------------------------------------------------------------------------------
```

`estimates store A`

The estimated log(odds) for this model are given by:

$$\log(\text{odds}) = -1.26 + 1.24 \times \texttt{agegrp} - 0.100 \times \texttt{agegrp}^2$$

The coefficient (log(odds ratio)) for **agegrp2** is less than zero, which means that the effect of age group tends to decrease with increasing age.

2. Omit **agegrp2**:

`logistic mf agegrp`

```
Log Likelihood = -729.23967
```

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   2.533891   .1615047    14.59   0.000     2.236321    2.871056
------------------------------------------------------------------------------
```

```
estimates store B
lrtest B A
```

```
Logistic:  likelihood-ratio test                  chi2(1)     =       2.32
                                                  Prob > chi2 =     0.1280
```

The likelihood ratio test indicates that the data are compatible with the null hypothesis (of linearity in the log(odds)) and that adding some non-linearity in does not improve the model from a statistical viewpoint.

8. Using continuous variables to test for interaction

In the previous session, we saw that in order to increase our power to detect interaction, we may use interaction terms with fewer categories than the main effects. Where the association between log(odds) of disease and an exposure is approximately linear, this may be done by looking for interactions with the quantitative variable.

Example:

```
. xi:logistic mf i.area*agegrp
i.area            _Iarea_0-1          (naturally coded; _Iarea_0 omitted)
i.area*agegrp     _IareXagegr_#       (coded as above)

Logit estimates                                   Number of obs   =       1302
                                                  LR chi2(3)      =     328.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -692.56131                       Pseudo R2       =     0.1919

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   1.616517    .429584     1.81   0.071     .9602327    2.721349
      agegrp |   2.223374   .2138813     8.31   0.000     1.841322    2.684697
_IareXageg~1 |   1.409155   .1895728     2.55   0.011     1.082548      1.8343
------------------------------------------------------------------------------
```

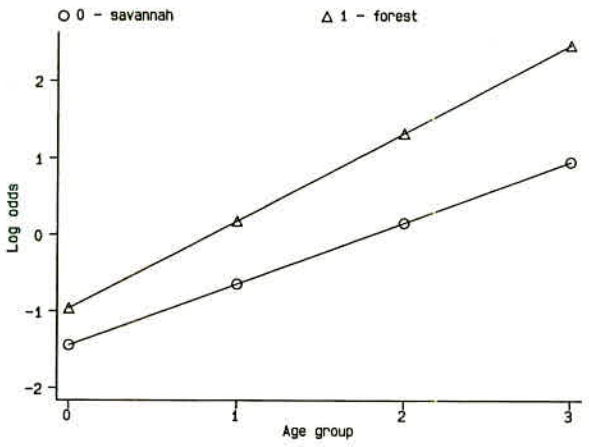
```
. logit
```

```
Logit estimates                                   Number of obs   =       1302
                                                  LR chi2(3)      =     328.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -692.56131                       Pseudo R2       =     0.1919

------------------------------------------------------------------------------
          mf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   .4802741   .2657466     1.81   0.071    -.0405796    1.001128
      agegrp |   .7990258   .0961968     8.31   0.000     .6104836     .987568
_IareXageg~1 |   .3429901   .1345294     2.55   0.011     .0793173    .6066629
       _cons |  -1.442564   .2061014    -7.00   0.000    -1.846515   -1.038612
------------------------------------------------------------------------------
```

The coefficients (log(odds ratios)) now have the following interpretations:

Coefficient	Value	Interpretation
`_Iarea_1`	0.4803	log(odds ratio) for area 1 versus area 0, in age group 0
`agegrp`	0.7990	Increase in log(odds) per increase in age group, in area 0
`_IareXageg~1`	0.3430	*Difference* between the increase in log(odds) per increase in age group in area 1 compared with area 0
`_cons`	-1.4426	log(odds) in area 0, age group 0

The fitted log(odds) based on this model are shown in the graph below.



The coefficient for **`_IareXageg~1`** represents the difference in slopes in the graph of log(odds) against age group for area 1 compared to area 0. If there is no interaction (coefficient=0) then the lines are parallel.

In this instance the likelihood ratio statistic is 6.46 (1df, p=0.011), indicating that there is evidence against the hypothesis of no interaction. The log(odds) of microfilarial infection increase more steeply with age in the forest than in the savannah area.

Note: when fitting such models in Stata the first term in the interaction must be the categorical variable. Fitting the model **`xi:logistic mf agegrp*i.area`** does not work.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 12 PRACTICAL

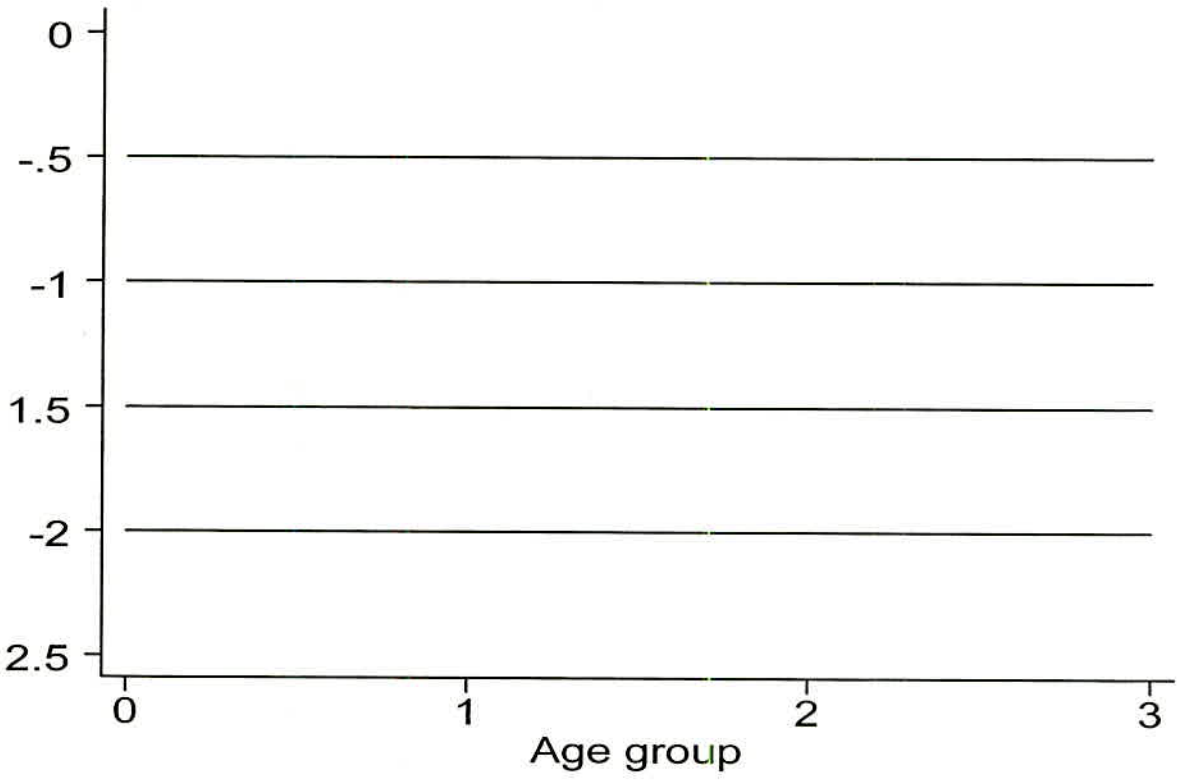
Objectives

By the end of this practical students will be able to:

(i) use logistic regression with quantitative exposures, and test whether fitting these exposures as quantitative rather than as categorical variables is appropriate;
(ii) interpret STATA output from a logistic regression model with a quantitative exposure.

1. Use data set ONCH667B.DTA, the prevalence study of onchocerciasis restricted to subjects aged 30 years or more (try `help oncho`). For this question the outcome variable is `lesions`, which indicates subjects who had eye lesions.

a) Use the `tab` command to derive the numbers of subjects who have eye lesions according to age group.
b) Use the `tabodds` command to calculate the odds of eye lesions in each group. Use a calculator or Stata to calculate the log(odds) of eye lesions in each age group. **Hint:** Stata will convert an odds of (eg) 0.13 into a log(odds) with the command `display log(0.13)`
c) Plot these log(odds) against age group.



d) Draw a straight line 'by eye' through the points you have plotted and estimate how much the log(odds) change for an increase in age of 20 years.

e) Fit a logistic regression model for eye lesions with the exposure being age group treated as a quantitative variable, obtaining your output on the log(odds) scale. **Make sure that you specified the correct outcome variable.**

f) What is the value of the slope of this regression line? What is the meaning of this slope? How does its value compare to your answer in part (c).

g) Fit a model with age group as a categorical variable and perform a likelihood ratio test of the null hypothesis that the effect of age group on the log(odds) of disease is linear versus the alternative that the relationship is more complicated than linear.

h) ***(optional — see supplementary material)***
Repeat the test of the null hypothesis that the effect of age group is linear, by creating a quadratic variable using

```
gen agegrp2=agegrp*agegrp
```

and fitting the appropriate logistic regression models. Use the `predict` command to calculate the fitted log odds for the model including `agegrp2`, and plot them on the graph in part (c).

2 a) Using data set ONCHALL, fit the model

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

If we treat age as a confounding variable, does our estimate of the effect of area differ when the confounding variable is treated as a linear effect rather than a categorical one? (compare this model with one where `agegrp` is categorical)

b) ***(optional — see supplementary material)***
Perform a likelihood ratio test for the interaction between `area` and `agegrp` treated as a linear effect.

3. Using the MWANZA data set, investigate the association between HIV infection and number of injections in the past year. Remember that the zero group should be excluded in order to confirm that a trend with increasing numbers is not simply induced by a difference between the zero category and the rest. See Clayton & Hills pp 256-258 for more on this.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 13

MATCHING IN CASE-CONTROL STUDIES

Objectives

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

(i) list the potential advantages of matching in case-control studies
(ii) list the potential disadvantages and limitations of matching in case-control studies
(iii) discuss the pros and cons of matching in specific situations

1. Introduction

In case-control studies, controls may be selected either by sampling at random from the 'non-case' population or they may be selected from sub-sets of the non-case population, matched to the cases with respect to particular characteristics that may be related to the risk of disease and/or exposure. Such matching may be performed at the individual level (e.g. siblings, neighbours) or at the group level; e.g. controls may be selected so that overall they have a similar age-sex distribution to the cases. Variables which might be used for matching include place of recruitment, place of residence, time of recruitment, age, sex, or other confounding factors.

In this session we consider the advantages and disadvantages of matched studies and examine the increase in efficiency that may be produced by matching.

2. The rationale for matching

Randomized, controlled trials are the 'gold standard' for establishing a causal link between an 'exposure' and disease. One reason for this is that randomization of sufficient numbers of subjects ensures that the 'exposed' and 'unexposed' groups are similar with regard to other potential risk factors for disease (e.g. age, sex). This similarity means that we can then ignore these other risk factors when comparing disease rates in the exposed and unexposed groups. (These risk factors are not associated with exposure and hence are not confounders). This applies both to risk factors of which we are aware and can measure and to risk factors about which we are unaware or are unable to measure.

Matching is another way of trying to ensure that our comparison groups are similar with regard to various factors. In a case-control study employing age-matching, for example, the investigator tries to ensure that the age distribution of the cases and controls is similar. Given that in an RCT (or cohort study) similarity of exposed and unexposed groups with regard to age (for example) means that age may be ignored in the analysis, it is tempting to believe that the same will be true for a matched case-control study. *This is not so*. Matching on age in a case-control study does not mean that age may be ignored in the analysis. This is illustrated below.

If matching in case-control studies does not simplify the analysis by enabling us to 'forget' about the matching variables, why match? There are three possible reasons for matching in case-control studies:

(i) to improve the efficiency/precision of the study,

(ii) to control confounding factors which are difficult to measure,

(iii) to provide a simple rule for the identification of controls.

Before considering (i) and (ii) in more detail we illustrate matching in a case-control study (and the analysis) with a hypothetical example.

2.1 An example of a matched case-control study

Consider a study of the association between exposure and disease, performed in a setting in which exposure is more common among men and disease is more common among women; i.e. sex is associated both with exposure and disease - it is a confounder. The tables below show the *population* from which we shall be sampling our cases and controls.

Males

	Exposed	Unexposed
Cases	450	10
Total population	900,000	100,000
Incidence/100,000	50	10

Females

	Exposed	Unexposed
Cases	200	360
Total	100,000	900,000
Incidence/100,000	200	40

In this population, the exposed have an incidence of disease 5 times that of the unexposed.

Suppose that a case-control study, frequency matched on sex, is conducted. Rather than selecting controls at random, controls are selected so that a similar number of controls as cases are female (without regard to the controls' exposure status). The results we would expect to obtain from such a study are shown in the tables below.

The numbers of controls in each category have been calculated as follows: there are 460 male controls in all to go with the 460 male cases; 90% of males are exposed; therefore we expect 414 exposed controls. Other expected entries in the table can be calculated in a similar fashion.

	Male			Female		
	Exposed	Unexposed	Total	Exposed	Unexposed	Total
Cases	450	10	460	200	360	560
Controls	414	46	460	56	504	560

For each stratum (males/females) we can calculate the odds ratio (=5.0 in both strata) and hence a summary estimate of the odds ratio adjusted for sex using the standard Mantel – Haenszel approach:

Summary odds ratio = 5.00 (95% c.i. 3.69, 6.79)

If we now perform an unstratified analysis, effectively ignoring sex (an 'unmatched' analysis), our results will be as shown below:

	Exposed	Unexposed	Total
Cases	650	370	1020
Controls	470	550	1020

OR = 2.06

The estimate of the odds ratio that we obtain from an unstratified analysis is biased. Rather than eliminate confounding, matching in a case-control study introduces a new confounding structure in place of the original structure. The effect of this new confounding, if we ignore it, is to lead us to *underestimate* the strength of the association (bias is always towards the null value 1.0 because by matching we have made cases and controls more similar than they would otherwise have been). This point is extremely important.

Matched design ==========> 'Matched' analysis

Methods for analysing individually matched case-control studies will be discussed in detail in another session.

3. Advantages of matching

- Matching **can** improve the precision of the odds ratio/increase the efficiency of the study.
- Matching **may** enable the investigator to control the effect of factors which cannot be easily measured (e.g. neighbourhood or sibling controls).
- Matching can provide simple, practical rules for identifying controls (e.g. neighbours)

3.1 Precision/efficiency gained by matching

It is generally the case that, for a study of a given size, the power of the study to detect a difference in exposure rates between cases and controls will be greatest if there are approximately equal numbers of cases and controls. For example, in the two situations depicted below, the odds ratio estimates (3.0) and the total numbers of persons included (200) in the study are identical but the widths of the confidence intervals are quite different because of the difference in the case:control ratio in the two studies.

Study 1 case:control ratio = 1:4

	Exposed	Unexposed	Total			
Cases	30	10	40	OR	=	3.0
Controls	80	80	160			(1.30,7.07)
	110	90	200			

Study 2 case:control ratio = 1:1

Cases	75	25	100	OR	=	3.0
Controls	50	50	100			(1.58,5.72)
	125	75	200			

When the analysis of a study involves stratification on the basis of some potentially confounding variable, the precision of the study will usually be maximal if the ratio of cases to controls is approximately the same in each stratum. Matching is a way of ensuring this balance.

We illustrate this effect by returning to our original example. As we saw, performing a matched study and a matched (stratified) analysis we obtain an unbiased estimate of the odds ratio (5.0) with a 95% confidence interval (3.69, 6.79).

Example (continued)

Suppose that we now conduct an *unmatched* study. We expect the following tables, this time with equal numbers of male and female controls.

	Male			Female		
	Exposed	Unexposed	Total	Exposed	Unexposed	Total
Cases	450	10	460	200	360	560
Controls	459	51	510	51	459	510

This time a stratified analysis gives the following results:

OR = 5.00 (95% c.i. 3.66, 6.85)

Our adjusted point estimate of the odds ratio is as for the matched study but the confidence interval around the point estimate is wider. Thus, matching on sex in this example improves

the precision of the study (or reduces the number of cases and controls required to obtain a given precision). However, we might note that the improvement in precision is not enormous.

3.2 Relative efficiency of matched and unmatched designs

Considerable research has been done to investigate the gains in statistical power associated with matching in case-control studies (either on a pair-wise basis or in strata) (e.g. Smith &Day. The design of case-control studies: the influence of confounding and interaction effects. Int J Epid 1984, **13**, 356-365; Schlesselman's book). Some general conclusions may be drawn.

1. Matching on a factor which is unrelated to disease or exposure does not result in substantial power loss unless there is a strong (OR=5+) association between exposure and disease. There is, of course, no possibility of *gain* in power under these circumstances and the matching procedure may complicate considerably the logistics of the study.

2. Matching on a variable which is associated with exposure but is not a risk factor for disease is a *bad strategy*, but unless the association with exposure is strong, it will not reduce power substantially.

3. In general it is only worthwhile matching on the basis of variables which are strongly confounding (e.g. substantial associations with disease and exposure).

4. If matching is part of the *design* then this must be taken account of in the *analysis*.

5. Attempting to match for more than a few variables is likely to be logistically inefficient, except in special situations.

3.3 Controlling confounders which are difficult to measure

It has been pointed out that, in addition to improving the precision of a study, *individual* matching of cases to controls *may* be equivalent to matching for a complex of underlying factors which may be only vaguely defined or difficult to quantify. For example, recruiting as controls siblings of cases may help the investigator to control certain genetic factors. Recruiting as controls neighbours of cases may help the investigator to control factors such as access to health services or socio-economic factors. The effectiveness of this strategy is not, however, guaranteed. Smith (*Int J Epidemiol*; **13**: 159-166, 1987) has cited the example of a case-control study to evaluate the effectiveness of BCG vaccination against tuberculosis. In this study neighbourhood controls were recruited with the idea that this would match controls to cases with regard to socio-economic status. When cases and controls were compared with regard to various socio-economic indicators, however, it was observed that cases tended to come from poorer households than controls.

Individual matching requires an appropriate stratified analysis. The strata consist of a single case together with their matched control(s). Analytical techniques appropriate for individually matched studies will be discussed in another session.

4. Disadvantages of matched studies

- cannot examine risk associated with matching factor (but can examine effect modification)
- data may be more difficult to present
- may be difficult to find suitable matches
- may have to interview cases first
- may have to interview many potential controls to find match
- multivariate analysis necessary for confounding variables that are not accounted for by the matching
- possibility of 'overmatching'

  - on variable associated with exposure but not disease power loss
  - on variable in causal pathway bias

"**To summarize**, the intricacies of matching in case-control studies and the relation of matching to confounding are much more complicated than one might at first suppose. Matching has often been employed when simpler and cheaper alternatives would have been preferable. Matching is clearly indicated only in sharply defined circumstances. In many study situations, the decision rests on cost and efficiency considerations which border on the imponderable." (Rothman, pp 249-50).

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 14

Analysis of Matched Case-Control Studies

1. Objectives

The objectives of this session are to become able to:

i. summarize the results of matched case-control studies in tabular form

ii. estimate the odds ratio

iii. test for effect modification (interaction)

iv. start to use conditional logistic regression for more complicated modelling.

2. Introduction

We have seen that confounding factors may be controlled in a case-control study by performing a stratified analysis. Stratification may also be incorporated into the design and conduct of a study by *matching*. In the previous session we saw that, for a fixed study size, a design and analysis including matching for one or more confounding factors is generally more efficient than a study that incorporates stratification only in the analysis. Matching may also enable an investigator to control confounding variables which cannot easily be quantified.

As we also saw in the previous session, stratifying (matching) in the design does not remove the need to stratify in the analysis. If we were to analyze a stratified (matched) case-control study ignoring the stratification, our estimate of the odds ratio is likely to be biased towards unity.

A frequency matched design, in which controls are selected in such a way as to ensure that they have (e.g.) a similar age distribution to that of the cases, can be analysed using the methods we have already met (Mantel Haenszel, logistic regression).

A more 'extreme' form of matching occurs when each control is individually matched to a particular case. These controls are chosen to be similar to the case with respect to one or more confounding factors. For example, if we wish to match on age and place of residence, for each case we might select as a control (or controls) a neighbour born within one year of the case. Individual matching in the design implies stratification in the analysis in which each stratum consists of a single case and their control(s). In this session we consider the analysis techniques appropriate

for the analysis of individually matched studies. More details can be found in chapters 5 and 7 of Breslow and Day, Volume 1 and chapters 19 and 29 of Clayton and Hills.

Example

To demonstrate the methods in this session we will use data from a matched case-control study conducted in southern Brazil to investigate potential risk factors for infant death from diarrhoea (Victora et al, 1987). Each infant dying from diarrhoea at less than 1 year of age (a case) was matched with 2 neighbourhood controls. Information on social and environmental factors, birth weight, and feeding mode were collected on 170 cases and 340 controls. To illustrate the methods for 1 control per case we have selected for each case the matched control closest in age to the case. In addition, we have matched on age by selecting only those case-control pairs where both the case and control were in the same age group (0-2, 3-5, ≥6 months). This reduced data set consists of 86 case-control pairs.

3. Matched pairs (1 control per case)

We begin by considering the situation where the exposure is a binary variable. In other words, it has only 2 levels, exposed and unexposed.

Suppose we have n matched pairs. Each pair can be thought of as a stratum. For each stratum (pair) there are four possible outcomes as follows:

	Exposure								Total for each table
	+	-	+	-	+	-	+	-	
Case	1	0	1	0	0	1	0	1	1
Control	1	0	0	1	1	0	0	1	1
	2	0	1	1	1	1	0	2	2
Total no. of pairs of each kind	n_{11}		n_{10}		n_{01}		n_{00}		n

In the notation used above, the first subscript for n represents the exposure status of the case and the second that of the control (0=unexposed, 1=exposed). Thus n_{01} is the number of pairs (tables) for which the case is unexposed (first subscript = 0) and the control is exposed (second subscript = 1).

The results of a pair-matched case-control study can therefore be presented in a table of the form:

		Control	
		Exposed	Unexposed
Case	Exposed	n_{11}	n_{10}
	Unexposed	n_{01}	n_{00}

Note that each cell in this table shows the number of case-control *pairs* of a particular kind. The usual table of numbers of individual cases and controls in each exposure category can easily be derived from the table of case-control pairs

	Exposure		Total
	+	-	
Case	$n_{11}+n_{10}$	$n_{00}+n_{01}$	n
Control	$n_{11}+n_{01}$	$n_{00}+n_{10}$	n
			2n

Exercise 1

Calculate the numbers of cases and controls 'exposed' and 'unexposed' for the following table drawn from the Brazilian study in which the 'exposure' is breast feeding.

Matched table

		Control	
	Feeding mode	Breast fed	Not breast fed
Case	Breast fed	24	6
	Not breast fed	29	27

Unmatched table

	Feeding mode	
	Breast fed	Not breast fed
Cases		
Controls		

3.1 Estimating the odds ratio

We can obtain an estimate of the odds ratio using the Mantel-Haenszel method for stratified data by applying the formula to each matched pair as follows (see Session 6):

$$\text{MHOR} = \frac{Q}{R} = \frac{\Sigma D_{1j}H_{0j}/N_j}{\Sigma D_{0j}H_{1j}/N_j}$$

In the top line, each of the n_{11} tables of the first kind contributes 1x0 (because $D_{1j}=1$ and $D_{0j}=0$). Each of the n_{10} tables of the second kind contributes 1x1, and so on. The bottom line can be considered similarly. Also, each N_j is two because each stratum is a pair. So we have:

$$\text{MHOR} = \frac{[(n_{11} \times 0) + (n_{10} \times 1) + (n_{01} \times 0) + (n_{00} \times 0)]/2}{[(n_{11} \times 0) + (n_{10} \times 0) + (n_{01} \times 1) + (n_{00} \times 0)]/2}$$

$$= \frac{n_{10}}{n_{01}}$$

Note that the concordant pairs (those for which the case and control have the same exposure) contribute nothing to the odds ratio estimate.

Exercise 2

Estimate the odds ratio of the association between breast feeding and infant diarrhoea mortality obtained from the Brazilian study (Exercise 1) using:

i. the matched table: $OR = n_{10}/n_{01} =$

ii. the unmatched table. OR =

3.2 Forming a confidence interval for the odds ratio

An approximate 95% confidence interval for the odds ratio may be calculated using one of the methods given in Session 6. There the error factor was given as:

$$EF = \exp(1.96 \times S) \qquad \text{where } S^2 = \frac{V}{QR}$$

Then note that:

$$Q = \frac{n_{10}}{2}$$

$$R = \frac{n_{01}}{2}$$

$$V = \frac{n_{10}}{4} + \frac{n_{01}}{4}$$

so $S^2 = 1/n_{10} + 1/n_{01}$

and $EF = \exp [1.96 \sqrt{(1/n_{10} + 1/n_{01})}]$

Note that, as for the point estimate of the odds ratio, the concordant pairs contribute nothing to the confidence interval.

The test-based method can also be applied.

Exercise 3

Calculate an approximate 95% confidence interval for the odds ratio of the association between breastfeeding and infant diarrhoea mortality.

Error factor =

95% CI =

When the number of discordant pairs ($n_{10} + n_{01}$) is small (e.g. less than 20), then an 'exact' 95% confidence interval based on the binomial distribution can be calculated (see e.g. Schlesselman, pages 211-212)

3.3 Test of the null hypothesis that the true odds ratio = 1

Like the estimate of the odds ratio and its confidence interval, the test of the null hypothesis that the true odds ratio is 1 is based only on the discordant pairs. When the true odds ratio is 1 there is no association between the exposure and the disease and therefore, given a discordant pair, it is equally likely to be of either type (probability of each type = 0.5). Given the total number of discordant pairs, the expected number of discordant pairs in which the case is exposed, $E(n_{10})$, is then $(n_{10}+n_{01})/2$. We require a test to assess whether n_{10} differs from its expected value more than would be expected by chance. When the number of discordant pairs is large (> 20 say) we can use the Normal approximation to the Binomial distribution to obtain an approximate test.

Under the null hypothesis that p=0.5, $var(n_{10}) = np(1-p) = (n_{10}+n_{01})/4$

Using the Normal approximation to the Binomial distribution gives:

$$\chi^2 = \frac{(n_{10} - E(n_{10}))^2}{} \quad \text{on 1 df}$$

$$\text{Var}(n_{10})$$

$$= \frac{(n_{10} - (n_{10}+n_{01})/2)^2}{(n_{10}+n_{01})/4}$$

$$= \frac{(n_{10}-n_{01})^2}{(n_{10}+n_{01})}$$

This test is usually known as McNemar's test for matched pairs. Note that some authors apply a continuity correction of -1 to the numerator. This approximate χ^2 test can also be derived by applying the Mantel-Haenszel χ^2 test, with the pairs as strata.

Exercise 4

Perform an approximate test of the null hypothesis that there is no association between infant feeding mode (breast fed vs non-breast fed).

Total number of discordant pairs = 29 + 6 = 35

χ^2 =

If the number of discordant pairs is small (≤ 20 say), an exact test based upon the Binomial distribution can be performed. This may be calculated using the table of cumulative probabilities of the binomial distribution with a value of $p = 0.5$ as the null hypothesis value (see (e.g.) Schlesselman, page 211).

3.4 Testing for heterogeneity of the odds ratio

While we cannot investigate a matching factor as a risk factor in its own right, we can test whether there is any evidence that a matching factor modifies the effect of (interacts with) the exposure of interest. To do this we define different levels of the matching factor (e.g. age groups) and look at the odds ratio estimate given by the pairs in each subgroup. The width of these subgroups may be broader than that used for the matching criteria so that these subgroups will be more heterogeneous than the individual matched pairs. For example, the pairs may be closely matched for age to within one year, but the subgroups may be defined by 10 year age groups. Considering only the discordant pairs we can present them in a table of the form:

Subgroup defined by levels of the matching factor

	1	2	3	i	I	Total
No of pairs with case exposed and control unexposed						
No of pairs with case unexposed and control exposed						

We thus have a 2 x I table, which can be analyzed using the χ^2 test for a 2xI table. The overall chi-squared test now indicates whether the odds ratio estimates vary according to the level of the matching factor. If the matching factor is on an ordinal scale then a test for trend can also be used.

Exercise 5

Low birthweight (< 3.0 kg) is crudely associated with about a 40% increase in infant mortality from diarrhoea (not significant, see table below).

		Control	
	Birthweight	<3.0	≥3.0 kg
Case	< 3.0	12	25
	≥ 3.0	18	31

OR = 25/18 = 1.39 (0.76,2.55)

We wish to see whether low birthweight has a greater effect on the risk of death from diarrhoea among younger infants than among older infants. Since the data are matched for age, we may stratify the pairs into three age groups as follows:

	Age					
	0-2 months		3-5 months		≥6 months	
	Control		Control		Control	
Birthweight	<3.0	≥3.0	<3.0	≥3.0	<3.0	≥3.0
<3.0	4	7	4	12	4	6
Case						
≥3.0	6	8	7	15	5	8

$OR_1 = 7/6 = 1.17$ $OR_2 = 12/7 = 1.71$ $OR_3 = 6/5 = 1.20$

To test for heterogeneity of these odds ratios we form a 2x3 table using the discordant pairs:

	Age group (months)		
	0-2	3-5	≥6
Case <3.0, control ≥3.0	7	12	6
Case ≥3.0, control <3.0	6	7	5

The chi-squared test of the null hypothesis of no association against a general alternative gives $\chi^2=0.35$ on 2 df, $p>0.5$. There is no evidence for a modifying effect of age on the odds ratio and it is reasonable to present an overall estimate for the odds ratio of 1.39 (=25/18) as calculated above. Since these data were matched for age and neighbourhood, this estimate of the odds ratio is adjusted for these two factors.

4. The analysis of studies recruiting more than 1 control per case

When more than one control per case is recruited the number of possible outcomes increases. For example, with two controls per case there are six possible outcomes for each triplet instead of the four possible outcomes when a single control per case is recruited. The methods above may be extended to situations when there are C controls per case or when there are a variable number of controls per case. The Mantel-Haenszel estimate of the odds ratio and test of significance can be obtained by considering each matched set as a separate stratum and using the methods for stratified data. The formula for the odds ratio in each situation reduces to:

$$\text{OR} = \frac{\text{Total no. of unexposed controls who have an exposed case}}{\text{Total no. of exposed controls who have an unexposed case}}$$

The formulae for these situations can be found in Breslow and Day Volume 1, pages 169-182. For example, if there are two controls per case, then there are six types of table, rather than the four in the paired case:

	exposure											
	+	-	+	-	+	-	+	-	+	-	+	-
case	1	0	1	0	1	0	0	1	0	1	0	1
control	2	0	1	1	0	2	2	0	1	1	0	2
		n_{12}		n_{11}		n_{10}		n_{02}		n_{01}		n_{00}

Doing the Mantel-Haenszel summation yields:

$$\text{MHOR}=(n_{11}+2n_{10})/(2n_{02}+n_{01}).$$

On the top line of this equation, the second type of table contributes 1 each, and the third type contributes 2 each. For each type, the number contributed is the number of unexposed controls when the table has an exposed case. The bottom line of the MHOR works out similarly. All the triplets with any discordancy contribute to the analysis. Only those triplets in which the case and both controls were unexposed (n_{00}), or in which all three were exposed (n_{12}) do not contribute.

Confidence intervals for the odds ratio in these more complicated situations can be calculated from the variance of the log odds ratio which is given in Rothman's book on page 273.

5. Analysis of matched case-control studies using STATA.

STATA can produce matched tables of studies with 1 control per case using the `match` command (an external command written by Michael Hills/David Clayton). The `match` command has the limitation that it will not produce correct tables for studies in which there is more than 1 control per case. The command, `mhodds`, provides an estimate of the odds ratio, an approximate confidence interval and performs an approximate test of the null hypothesis of no association, for any number of controls per case, providing the data set contains a variable indicating to which pair each case and control belong. `mhodds` works by stratifying on this pairing variable and calculating the Mantel-Haenszel estimate of the odds ratio. It does not produce tables.

6. Exposures with more than 2 levels and further analyses.

As we have seen, the analysis of an individually matched case-control study requires the formation of a table of the exposure of each case according to the exposure of its matched control or controls. Suppose that each case has just 1 control but that the exposure has 3 levels. The table that we have to form to analyze this would be a 3x3 table. Even in the simplest situation the analysis is quite involved. In this situation one could restrict attention to comparing only two levels of the exposure at a time and use the methods in sections 2 and 3. This would give odds ratios for the comparison of different pairs of levels of the exposure, but these estimates may not be consistent. The odds ratio for comparing levels 1 and 3 will not in general equal the product of the odds ratio for comparing levels 1 and 2 with the odds ratio for comparing levels 2 and 3.

Other methods of analysis using statistical modelling techniques are therefore recommended for all situations when the exposure has more than two levels.

7. Allowing for other factors

Often one may need to take into account confounding and effect modifying factors on which the cases and controls have not been matched. For instance, we might wish to control for a potential confounder by stratification. This is not possible, because the data are already stratified into pairs of cases and controls so that no further stratification is possible. These analyses cannot be performed easily or efficiently using classical methods. The only way to investigate confounding or effect modification using classical methods is to restrict attention to those case-control pairs that are homogeneous with respect to the confounding or effect modifying factor of interest, so that much of the data may have to be discarded. Thus for these analyses we need to use statistical modelling techniques.

8. Conditional Logistic Regression

We have seen that for a stratified design, ignoring stratification in the analysis gives biased estimates. This causes problems if we try to use ordinary logistic regression to analyze individually matched case-control studies. We might be tempted to try to analyze a study with one matched control for each case using ordinary logistic regression and allowing for the matching by including the number of the matched set in the model as a categorical variable. There are two problems when we do this, however:

1. The number of terms in the model becomes very large (the number of matched sets is half the number of individuals in the study).

2. The estimate of the odds ratio is biased. This is related to the first point: if the number of parameters in a model is comparable to the number of observations, then maximum likelihood estimators are not necessarily unbiased (Breslow and Day, 1980, volume 1, p249; Cox and Hinkley, 1974, p292). It can be shown that for one control per case, the estimated odds ratio is exactly the square of the correct value. The magnitude of the bias goes down as the number of individuals in each stratum increases.

We use instead *conditional logistic regression* to analyze matched and other finely stratified studies. Conditional logistic regression is a modification of (unconditional) logistic regression in which the likelihood is rewritten to take account of the matching. In effect, parameter estimates are obtained by looking at the values of the explanatory variables among cases and controls *within each stratum.*

Conditional logistic regression allows the same flexibility in modelling matched case-control studies as does logistic regression for cross-sectional and unmatched case-control studies.

- Each stratum may contain any number of cases and controls. The only limitation is that for strata containing large numbers of cases and controls the amount of computation becomes large, but this is less of a problem with recent increases in computing power.

- We can allow for potential confounders which were not matched on by including them as factors in the model.

- We can model the effects of exposures with two or more levels, either as factors or as continuous variables.

8.1 Conditional logistic regression using STATA

The `clogit` command is used in STATA to perform conditional logistic regression. The command is very similar to the `logit` command for (unconditional) logistic regression, except that it requires an additional piece of information indicating to which matched set each individual belongs. This is supplied in the `strata()` option (`str` for short). This can be done most conveniently by including in the file a variable (e.g. `set`) which gives the number of the matched set for each case and control. The outcome variable (e.g. `case`) should be coded as 1 for cases and 0 for controls. Then a typical `clogit` command would look like this:

```
clogit case age bf, str(set)
```

If the explanatory variables `age` and `bf` are to be treated as factors (categorical) then the command would look like this:

```
xi: clogit case i.age i.bf, str(set)
```

Conditional logistic regression will be covered in more detail in the Study Unit 'Advanced Statistical Methods in Epidemiology'. A clear description is also given by Collett in his 1991 book on analysis of binary data.

References

Collett D (1991). *Modelling Binary Data*. London, Chapman and Hall.

Cox DR, Hinkley DV (1974). *Theoretical Statistics*. London, Chapman and Hall.

Victora CG, Smith PG, Vaughan JP, Nobre LC, Lombardi C, Teixeira AM, Fuchs SM, Moreira LB, Gigante LP, Barros FC (1987). Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil. *Lancet* **2**(8554): 319-322.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 14 PRACTICAL

Analysis of Matched Case-Control Studies

The objectives of this practical are to:

become able to use the `match` command in STATA to analyse data from matched case-control studies, where there is one control per case

become able to use the `mhodds` and `clogit` commands to analyse data from individually matched case-control studies

see that an unmatched analysis of both a) individually matched and b) frequency matched case-control studies can bias the estimates of effect (ORs)

Individually matched case-control studies

The practical uses two datasets, DIABRAZ.DTA (1 control per case) and DIABRAZ2.DTA (2 controls per case) and three commands within STATA: `match`, `mhodds` and `clogit`. See the course handbook for details of the datasets. If you have not used the STATA commands before, use the online help facility to check the syntax of these commands.

1. Using DIABRAZ.DTA and the `match` command reproduce the 2x2 matched table of the association between breast feeding (variable `bf`) and diarrhoea mortality. Using the `mhodds` command, estimate the odds ratio, calculate a confidence interval for the odds ratio and test the null hypothesis of no association. Compare your answers with those given during the lecture. Are they the same?

2. Perform similar analyses for two more variables, water supply (`wat2`) and birthweight (`bwtgp`). From the 2x2 matched tables, construct unmatched tables and calculate unmatched estimates of the odds ratios. What is the effect of ignoring the matching in the analysis? Is it the same in both cases? If not, why not?

3. In order to see whether there was any evidence that the odds ratio for breast feeding is different in very young infants from older children, the data were stratified into two age groups (0- 2, ≥3 months).

		Ages 0-2 months		Ages ≥ 3 months	
		Control		Control	
	Breast fed	Yes	No	Yes	No
Case	Yes	11	2	13	4
	No	9	3	20	24

Estimate the odds ratio in each age group and test the null hypothesis that the odds ratio is the same in the two age groups. You may find the `cci` command in STATA useful.

4. The table below shows the effect of breast feeding in the whole study (170 cases and 340 controls, DIABRAZ2.DTA).

		Number of controls breast fed		
		0	1	2
Case	Breast fed	5	24	21
	Not breast fed	18	66	36

Use `mhodds` to estimate the odds ratio for breast feeding versus no breast feeding, test the null hypothesis that the true odds ratio is 1 and calculate a 95% confidence interval for the estimated odds ratio.

Check that you can derive the same odds ratio manually using methods described in the lecture.

5. Using the `clogit` command, repeat the matched analyses in questions 1, 2 and 4.

Frequency matched case-control studies

6 (a) For the following hypothetical population, draw up tables, stratifying on sex, for the results you would expect to obtain from an *unmatched* case-control study conducted in this population.

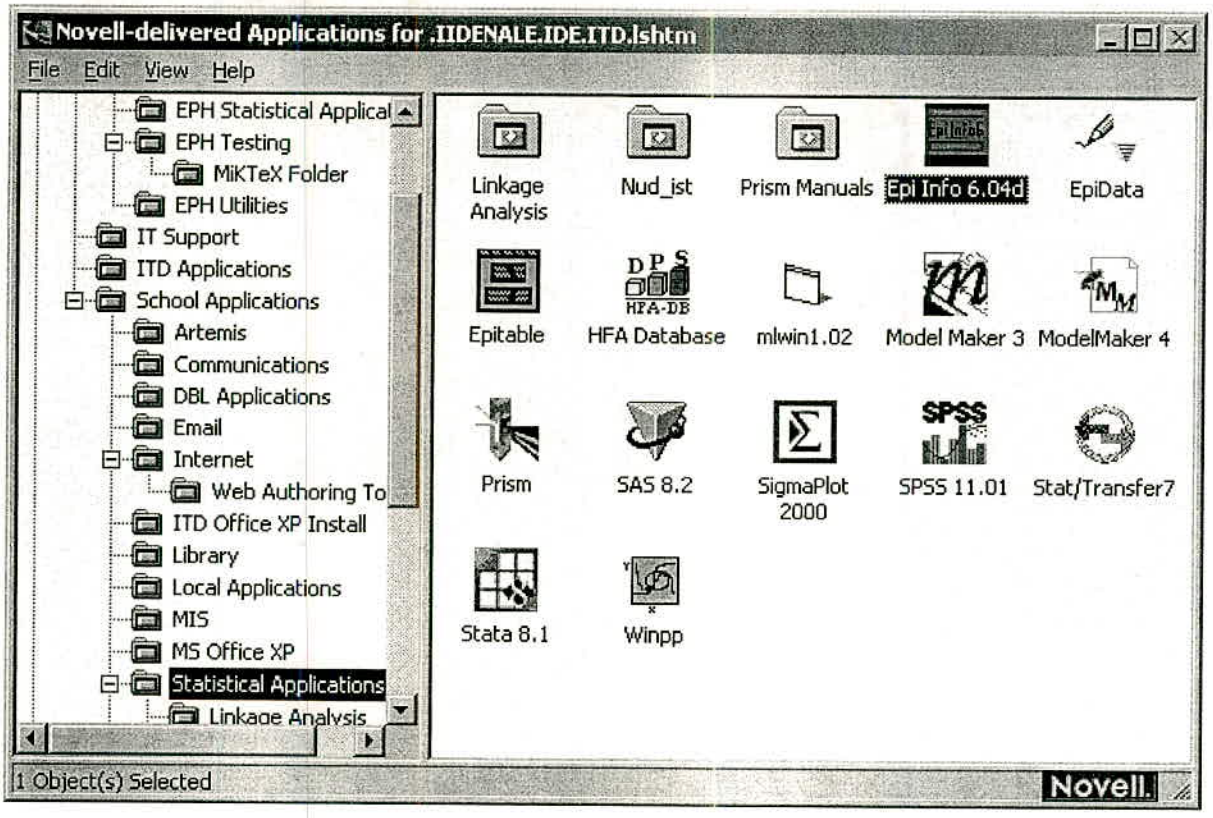
Males

	Exposed	Unexposed
Cases	450	10
Total population	900,000	100,000

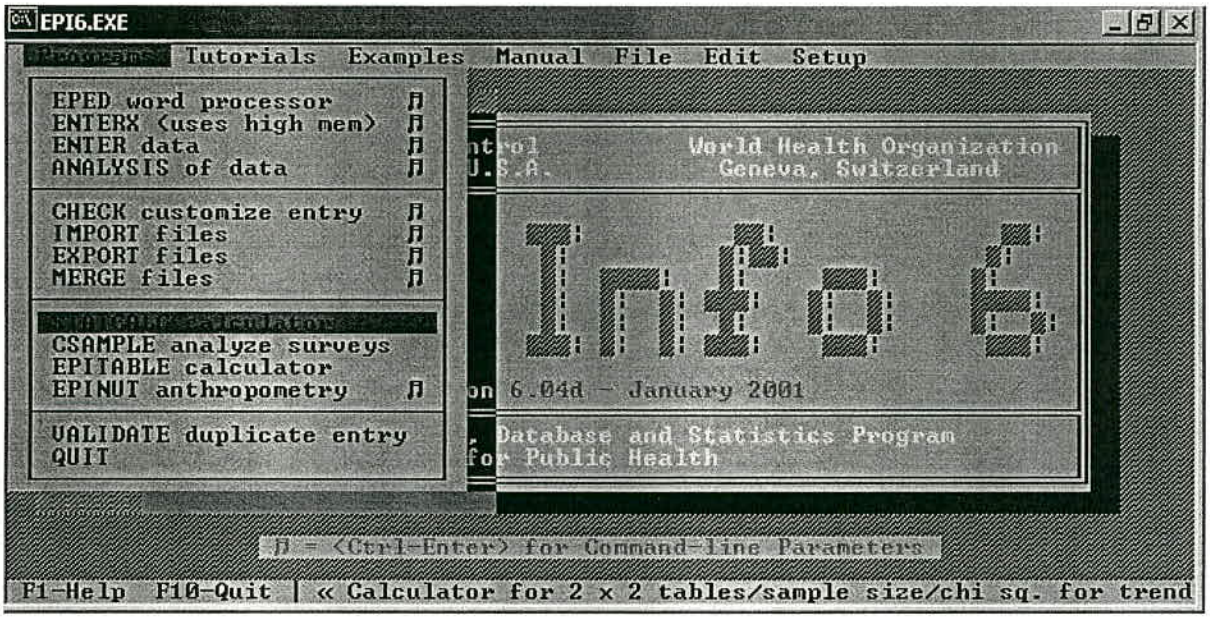
Females

	Exposed	Unexposed
Cases	50	90
Total	100,000	900,000

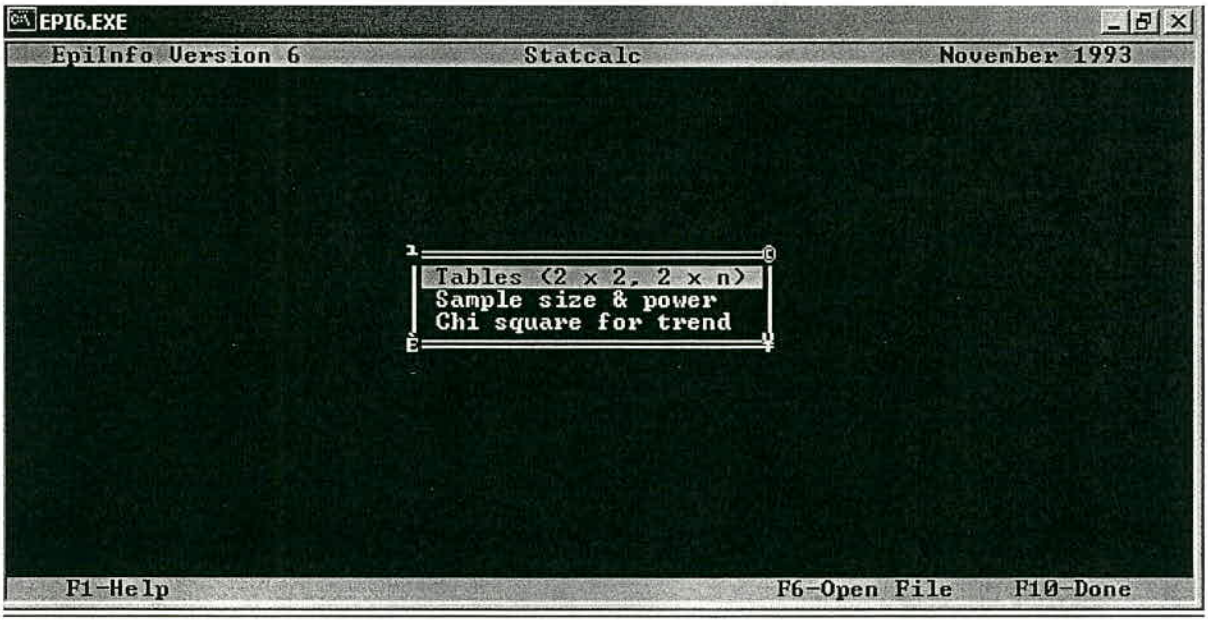
Use the STATCALC program of Epi Info to obtain a summary odds ratio and a confidence interval for this OR. This program is available in the 'Statistical applications' of the 'Novell-delivered applications' window:



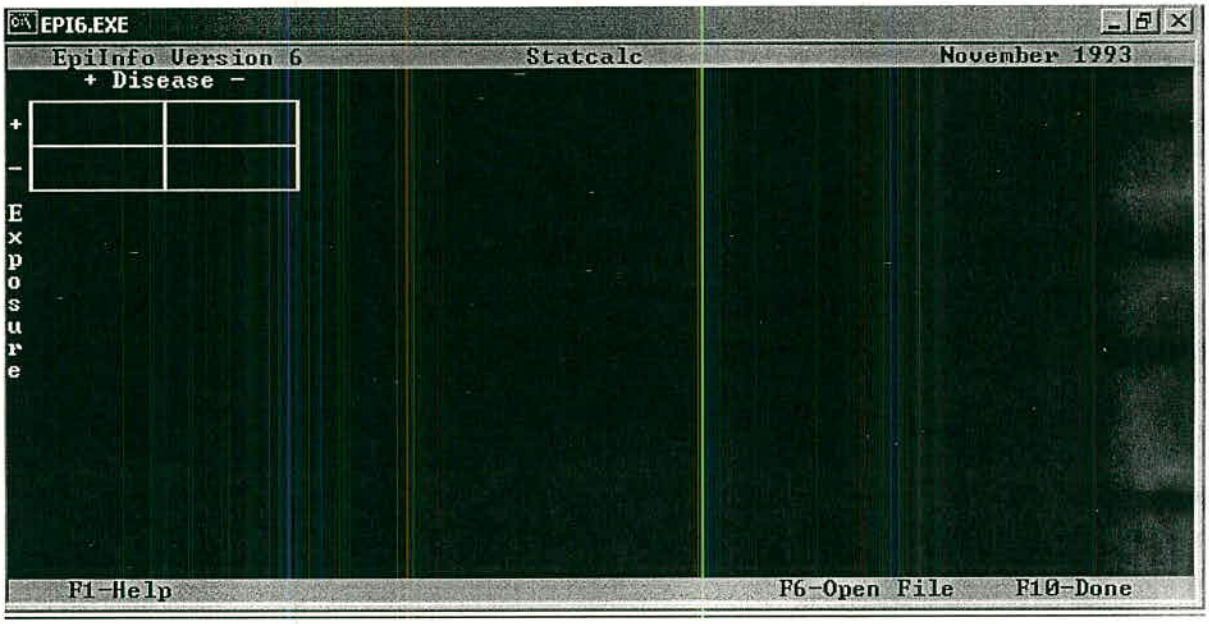
Double-click on the Epi Info icon and choose STATCALC from the 'Programs' menu:



Then choose the first option ('Tables (2 x 2, 2 x n)'):



Then fill in the numbers for the first stratum (eg males) of your case control study (note the orientation of the table; 1st column cases, 2nd column controls). Press ENTER after each of the four cell frequencies, then F4 (or ENTER again) to show the results for that stratum.



After the stratum results have been shown, press F2 to enter the second stratum (eg females), and proceed as for the first one.

After seeing the results for the second stratum, press ENTER again to indicate that there are no more strata. Then you will get the summary results, including the Mantel-Haenszel odds ratio and confidence interval.

(b) Now draw up tables for the results that you would expect to obtain from a case-control study *frequency matched* (not individually matched) on sex. Again, use STATCALC within Epi Info to obtain a summary odds ratio and a confidence interval for this OR. Calculate a confidence interval for the matched study. What is the effect of matching in this situation? Why? What happens if you perform unstratified analyses of each study?

Optional - Individually matched studies

7. Investigate whether social class, income or maternal education confound the associations between diarrhoea mortality and breast feeding, birthweight or water supply, using the data set DIABRAZ2.DTA

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 15

Strategies of analysis

Objectives

By the end of the session students will:

(i) know that classical Mantel-Haenszel methods should be used before moving to regression methods
(ii) be able to identify the types of variables which should be included in a regression model
(iii) be able to formulate a sensible modelling strategy to build a regression model

1. Introduction

This session focusses on the practicalities of statistical analysis using the range of techniques covered in this course: specifically stratification and regression modelling. We consider when it is appropriate to apply stratification methods, given the knowledge that regression models can often be used to obtain similar results. We consider how to formulate a modelling strategy in regression modelling, putting together the techniques learnt over the previous sessions.

2. Univariate analysis

As we noted in previous sessions, simple, univariate analyses are the building blocks of more sophisticated techniques, and should precede both stratification and regression modelling. Scrutinising univariate tables and graphs should inform your decisions about how to group each variable and the likely importance of each variable in the later analysis.

3. Allowing for confounders — classical stratification versus regression models

Confounding is present to some extent in all observational studies, and some attempt must be made to remove confounding effects in the analysis, for each of the main exposures of interest. We have two tools available for this task: the classical (Mantel-Haenszel) methods based on stratification, and regression modelling. Each has advantages. Reasons for using classical methods are:

i) The availability of the data in simple tables keeps the investigator in touch with the data. In contrast, regression modelling is a 'black box' approach i.e. the

mathematical calculations are complex and largely hidden so there is a tendency to simply believe the output. It is therefore easier to make disastrous errors in regression modelling.

ii) Regression models generally involve additional assumptions; for example, that certain interactions can be omitted.

We recommend that classical methods should *always* be used in the initial phase of the analysis. If there are three or fewer potential confounders then classical methods may be sufficient for the entire analysis. In this case, stratify on each variable in turn, and then on the cross-classification of these variables, and see how the effect estimate is altered.

The problem is that a large number of 'potential confounders' are usually measured, and any attempt to control simultaneously for all of them, by stratifying on the cross-classification of all the confounders, quickly gets into trouble because of the very large number of strata and the very small number of cases in each stratum. If too many strata are defined, the result will be rapidly widening confidence intervals for the effect estimate.

So where the number of potential confounders is too large to be controlled satisfactorily by stratification alone, regression modelling (assuming no or only limited interaction between confounding variables) is more useful.

Despite this, classical methods can often, if applied sensibly, provide reasonable estimates of effect from which most of the bias due to the measured confounding factors has been removed. (The exception is analysis of matched case-control studies, for which classical techniques are of more limited use).

We have seen that regression methods for the effect of two categorical variables involve exactly the same stratification assumptions, and hence give essentially the same results, as classical methods. Regression models have the following advantages over classical methods:

i) By assuming that there is no interaction between confounders, we can greatly reduce the number of strata (the number of parameters used to describe the effect of the confounders).

ii) The effect of each variable, controlled for the effect of the others, can be seen.

iii) Dose-response effects can be examined more flexibly.

4. Modelling strategy in regression analysis

This topic will be covered in an interactive session during the lecture. Further notes based on the overheads of the lecture will be given at the end of the session. You can use the space below to make your own notes from this session.

References

The following papers give ideas for modelling strategies based on thinking about causal mechanisms and how different variables inter-relate:

The role of conceptual frameworks in epidemiological analysis: a hierarchical approach. Victora CG, Huttly SR, Fuchs SC, Olinto MT. *Int J Epidemiol* 1997, 26: 224-7.

Data, design, and background knowledge in etiologic inference. Robins JM. *Epidemiol* 2001, 11: 313-20.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 16

PRESENTATION OF STATISTICAL RESULTS

Aim: To improve students' ability to write clear reports of statistical analyses of epidemiological data.

Introduction

In your professional lives you are likely to have to prepare reports of statistical analyses. More immediately, you must prepare such a report as the assignment for this Study Unit. To do this, you need the skills to:

1. **Select** from statistical analyses carried out **key** results for presentation in a **limited** space;
2. Provide results **adequate** to support interpretation but **not so detailed** as to distract from focus on key findings;
3. Choose appropriately between **tables, graphs, or text** as a means of presenting results;
4. Construct useful, **clear**, and clearly-labelled **tables** and graphs;
5. Summarise the role of chance, reporting **confidence intervals or p-values** as appropriate;
6. Make results **accessible** to readers who do not have detailed knowledge of the statistical methods used;
7. Recognise limitations in presentations of statistical results, and **advise** as to how they could be **improved**.

This session is designed to help you increase these report-writing skills, in particular 3, 4, and 7. For guidance you might like to look at a statistical text-book, most of which have sensible advice on the presentation of statistical results, and some of which have entire chapters on this.

One fairly full treatment concerned with presenting simple results is given in :

Bland M. An introduction to Medical Statistics OUP Oxford 1995. Chapter 5.

Activities

The two papers (one on onchocerciasis and epilepsy, one on HIV and mortality) should be read in advance of the session. They should be read as if invited to be a co-author of the paper, sent a draft manuscript for comment and revision prior to submitting it to a journal.

9:30-10:45 In practical group (dividing into groups of 4-5)

In groups, discuss whether you think that the manuscripts are clear, and if not in what ways they are deficient. You are to check the presentation of statistical methods and results, and also discuss strategies to improve presentation.

You may find the following check-list helpful:

1. Has enough information been presented on the design of the study and the statistical methods used for readers to interpret properly the results?
- if not, what more information should be presented?
- is redundant information presented?

2. Has your co-author chosen a logical order in which to present the results?
- how would you structure the results section?

3. Are there components of the results section which are lacking?
- what additional data/analyses/results would you suggest including?

4. Are there any results which are not important in the context of the paper and could be omitted?
- which results would you have omitted?

5. Which results, if any, would you propose presenting in tables or figures?
- draw up tables/figures to show how you would present the data/results. Fill in values where they are available, and leave blanks where values are not available from the article but would be by further analyses of the data.

6. Has your co-author discussed appropriately the limitations of the data?
- if not, what points have they missed?

10.45-11:10 Each group should prepare a presentation on one of the 2 papers (at most 4 OHPs)

11:10-11.30 Break

11:30-12:30 Plenary session

Each group will present (5-7 minutes) their proposals to improve the manuscript, receive comments, and make comments on other groups' outlines.

At the end of the session the practical group leader will make some general comments.

Statistical Methods in Epidemiology (2402)

Solutions

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 1 SOLUTIONS

1.

```
. use whitehal
```

(Whitehall Study - 10% sample)

```
. tab all
```

death from all causes	Freq.	Percent	Cum.
0	1,274	75.97	75.97
1	403	24.03	100.00
Total	1,677	100.00	

There were 403 deaths from all causes

```
. tab chd
```

death from chd	Freq.	Percent	Cum.
0	1,523	90.82	90.82
1	154	9.18	100.00
Total	1,677	100.00	

There were 154 deaths from coronary heart disease

```
. tab smok
```

never/ex/1-14/15-24/25+	Freq.	Percent	Cum.
1	317	18.90	18.90
2	646	38.52	57.42
3	310	18.49	75.91
4	279	16.64	92.55
5	125	7.45	100.00
Total	1,677	100.00	

The frequency column indicates the number of individuals in each smoking category.

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

```
never/ex/1 |
-14/15-24/ |          smok2
       25+ |           0           1|     Total
-----------+----------------------+----------
         1 |         317           0 |       317
         2 |         646           0 |       646
         3 |           0         310 |       310
         4 |           0         279 |       279
         5 |           0         125 |       125
-----------+----------------------+----------
     Total |         963         714 |     1,677
```

This is an important check that you have created the new variable correctly.

```
. tab all smok2, col
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+
```

```
death from |          smok2
all causes |           0           1 |     Total
-----------+----------------------+----------
         0 |         795         479 |     1,274
           |       82.55       67.09 |     75.97
-----------+----------------------+----------
         1 |         168         235 |       403
           |       17.45       32.91 |     24.03
-----------+----------------------+----------
     Total |         963         714 |     1,677
           |      100.00      100.00 |    100.00
```

Risk among never-smokers = 17.45% i.e. 17.45 per 100

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

```
           | Summary of death from all causes
     smok2 |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 |   .17445483   .37969731         963
         1 |   .32913165   .47022728         714
-----------+------------------------------------
     Total |   .24031008   .42739919        1677
```

Risk among never-smokers (category 1) = 0.174 x 100 = 17.4% as before

```
. cs all smok2
```

```
                 | smok2
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |       235         168  |       403
        Noncases |       479         795  |      1274
-----------------+------------------------+-----------
           Total |       714         963  |      1677
                 |                        |
            Risk |  .3291317    .1744548  |  .2403101
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .1546768       |   .112695    .1966586
      Risk ratio |          1.88663       |  1.587309    2.242394
 Attr. frac. ex. |         .4699543       |  .3700029     .554048
 Attr. frac. pop |         .2740428       |
                 +-----------------------------------------------
                           chi2(1) =    53.73  Pr>chi2 = 0.0000
```

The risks are 0.329/person and 0.174/person for smokers and non-smokers respectively and the risk ratio = 1.89 (95% CI 1.59-2.24).

2.

```
. stset timeout, fail(all) origin(timein) id(id) scale(365.25)
                id:  id
     failure event:  all != 0 & all < .
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
      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
```

Check that there are 1677 observations and 403 failures (deaths from all causes).

```
. strate smok2
      failure _d:  all
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)
```

```
+-------------------------------------------------------------+
| smok2      D          Y          Rate       Lower       Upper |
|-------------------------------------------------------------|
|     0    168     1.6e+04     0.0102673   0.0088264   0.0119434 |
|     1    235     1.1e+04     0.0209024   0.0183937   0.0237532 |
+-------------------------------------------------------------+
```

The rate for the combined group of non-smokers and ex-smokers is 0.010267 per person-year (168/16363) or (1000 x 0.010267) i.e. 10.267 per 1000 person-years. The rate for smokers is 20.902 per 1000 years. The rates per 1000 person-years are given directly below

```
. 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)

  +-----------------------------------------------------+
  | smok2      D          Y       Rate      Lower      Upper |
  |-----------------------------------------------------|
  |     0    168    16.3626    10.2673     8.8264    11.9434 |
  |     1    235    11.2427    20.9024    18.3937    23.7532 |
  +-----------------------------------------------------+

      . strate smok, 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)

  +-----------------------------------------------------+
  | smok       D          Y       Rate      Lower      Upper |
  |-----------------------------------------------------|
  |     1     33     5.5921     5.9012     4.1953     8.3007 |
  |     2    135    10.7706    12.5342    10.5885    14.8373 |
  |     3     89     5.0148    17.7473    14.4180    21.8454 |
  |     4     98     4.2911    22.8382    18.7360    27.8385 |
  |     5     48     1.9368    24.7826    18.6761    32.8857 |
  +-----------------------------------------------------+
```

This time we have more detail with separate rates for non-smokers and ex-smokers and separate rates for sub-categories of smoker.

3.

```
      . stmh smok2

        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% conficence limits

------------------------------------------------------------
     RR       chi2       P>chi2     [95% Conf. Interval]
------------------------------------------------------------
  2.036      51.63       0.0000         1.670       2.482
------------------------------------------------------------
```

From this we can conclude that smokers experience twice the death rate (2.036) of non-smokers with 95% confidence interval 1.670-2.482. Note that the category with the lower code value (code 0 = non/ex smokers) is taken as the reference category unless you specify otherwise. The estimated rate ratio is slightly larger than the estimated risk ratio (1.89).

4.

```
use mwanza, clear

tab case ed, row chi
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

Case/control                Education
                1          2          3          4 |     Total
----------+--------------------------------------+----------
        0 |      263         51        255          5 |       574
          |    45.82       8.89      44.43       0.87 |    100.00
----------+--------------------------------------+----------
        1 |       49         24        110          6 |       189
          |    25.93      12.70      58.20       3.17 |    100.00
----------+--------------------------------------+----------
    Total |      312         75        365         11 |       763
          |    40.89       9.83      47.84       1.44 |    100.00

      Pearson chi2(3) =  26.7371   Pr = 0.000
```

There is strong evidence against the null hypothesis of no association between education and HIV infection ($p<0.001$). Those with no education seem to be under-represented among cases relative to controls (25% against 46%). Also note that the group with the highest education (code 4) is small so that any estimates of effect for this group will have wide confidence intervals.

As this is a case-control study and the cases and controls 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 dealt with in more detail in a later session.

```
. gen ed2 = ed
. recode ed2 2/4 = 2
(376 changes made)
```

```
. tab ed ed2
           |          ed2
 Education |         1          2 |     Total
-----------+----------------------+----------
         1 |       312          0 |       312
         2 |         0         75 |        75
         3 |         0        365 |       365
         4 |         0         11 |        11
-----------+----------------------+----------
     Total |       312        451 |       763
```

Don't forget to check you have created the variable correctly.

```
. tab case ed2, row chi
Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |       263        311 |       574
           |     45.82      54.18 |    100.00
-----------+----------------------+----------
         1 |        49        140 |       189
           |     25.93      74.07 |    100.00
-----------+----------------------+----------
     Total |       312        451 |       763
           |     40.89      59.11 |    100.00

          Pearson chi2(1) =  23.2789   Pr = 0.000
```

```
. mhodds case ed2

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

  ------------------------------------------------------------------------
   Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
  ------------------------------------------------------------------------
     2.416169      23.25        0.0000         1.668360   3.499168
  ------------------------------------------------------------------------
```

This indicates that the odds of HIV infection in the educated women is 2.42 times that in the uneducated women, with confidence interval 1.67-3.50. It is calculated from the 2x2 table as (140 x 263)/(49 x 311).

5.

```
. tab skin

      Skin |
  incisions |
  or tatoos |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        342       44.82       44.82
          2 |        420       55.05       99.87
          9 |          1        0.13      100.00
------------+-----------------------------------
      Total |        763      100.00
```

One individual has `skin` coded as 9. This represents a missing value.

```
. mhodds case skin

Score test for trend of odds with skin

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

------------------------------------------------------------------
Odds ratio    chi2(1)       P>chi2       [95% Conf. Interval]
------------------------------------------------------------------
  1.191688       1.40       0.2368        0.891231   1.593439
------------------------------------------------------------------

. mhodds case skin, c(2,1)

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

------------------------------------------------------------------
  Odds ratio    chi2(1)       P>chi2       [95% Conf. Interval]
------------------------------------------------------------------
   1.326275       2.74       0.0977        0.948473   1.854564
------------------------------------------------------------------
```

The answers are different because there is a missing value for skin which is coded 9 and if there are more than two categories the **mhodds** command yields an estimate of the average increase in odds from one category to the next. The first analysis has treated the missing value as though 9 was a valid code while the second analysis has excluded it. You can exclude the missing value and treat the variable as a binary variable for most analyses by recoding the missing value to system-missing (.) which STATA will ignore for many of its commands.

```
. recode skin 9=.
(skin: 1 changes made)
. mhodds case skin
Maximum likelihood estimate of the odds ratio
Comparing skin==2 vs skin==1

------------------------------------------------------------------
  Odds ratio    chi2(1)       P>chi2       [95% Conf. Interval]
------------------------------------------------------------------
   1.326275       2.74       0.0977        0.948473   1.854564
------------------------------------------------------------------
```

Key Points

- When analysing cohort studies we can use risks and risk ratios or rates and rate ratios. Rates are often preferred to risks because they allow for varying follow-up times and take account of when events occur.

- The analysis of cross-sectional and case control studies is often based on odds ratios. If the outcome is rare the odds ratio, risk ratio and rate ratio will all be the same (as near as makes no difference). If the outcome is not rare the odds ratio will be further from 1 than the risk ratio and care needs to be taken in its interpretation.

- In STATA we can use
  - the `tab` command to display risks, for example in a cross-sectional study. The `cs` command can be used if both exposure and outcome are coded as 0/1 variables.
  - the `stset` command followed by `strate, stmh` to display rates and calculate rate ratios in a cohort study;
  - the estimation of odds of disease is usually only possible for cross-sectional (or fixed cohort) studies, but not for case-control studies;
  - odds ratios can be obtained for case-control and cross-sectional studies with the `mhodds` command.

It is important to be familiar with your data set before you do any modelling. Missing and peculiar values can distort your results.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 2

SOLUTIONS

*Q1

```
.. log using sme2
 . cd h:\sme
 . use whitehal,clear
```

*Q2

The outcome is the overall mortality **all**, while 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.

```
. stset timeout,fail(all) origin(timein) id(id) scale(365.25)
                id:  id
     failure event:  all ~= 0 & all ~= .
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
      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 data set. The follow-up time starts at entry, hence the minimum is zero, and lasts up to the maximum of 19.4 years.

*Q3

The variable **agein** holds age at entry into the study. To categorise it into groups (40-44, 45-49, 50-54, ..., 65-69) we use the command **egen** with the **cut** option. Adding the option **label** makes the definition of each new category to be shown in all analyses involving **agecat**.

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

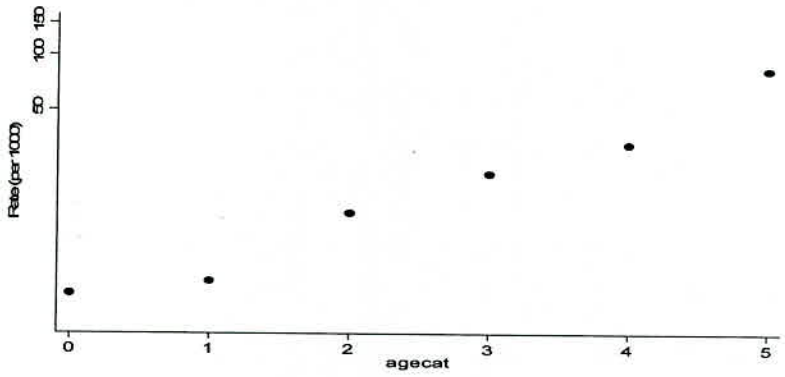
To examine how mortality rates change with age at entry we use **strate** and obtain:

```
. strate agecat, per(1000)
Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)
   agecat      _D          _Y       _Rate      _Lower      _Upper
      40-      24     4.91862       4.879       3.271       7.280
      45-      45     7.85489       5.729       4.277       7.673
      50-      82     6.05975      13.532      10.898      16.802
      55-     118     5.28201      22.340      18.652      26.757
      60-     101     3.09485      32.635      26.852      39.662
      65-      33     0.39525      83.492      59.357     117.441
```

*Q4

The rates increase with age at entry. This can be seen graphically when the option **graph** is added to the **strate** command.

If the RRs between successive categories are similar, the differences between successive log (RR) should be constant. Hence plotting the RRs on a log scale should show a linear relationship with age. This can be done by adding the option **yscale(log)** to the command **strate**.



This plot shows that the rates increase fairly linearly with **agecat** (on a log-scale).

*Q5

To use **stmh** with **agecat** we need first to check how this new variable is coded and then select the two categories we wish to compare.

```
. tab agecat, nolabel
    agecat |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |        277       16.52       16.52
         1 |        445       26.54       43.05
         2 |        362       21.59       64.64
         3 |        340       20.27       84.91
         4 |        215       12.82       97.73
         5 |         38        2.27      100.00
-----------+-----------------------------------
     Total |       1677      100.00
. stmh agecat, c(1,0)
Maximum likelihood estimate of the rate ratio
  comparing agecat==1 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits
------------------------------------------------------------
        RR       chi2       P>chi2     [95% Conf. Interval]
------------------------------------------------------------
     1.174       0.40       0.5250        0.715       1.927
------------------------------------------------------------
```

```
. stmh agecat, c(2,0)
Maximum likelihood estimate of the rate ratio
  comparing agecat==2 vs. agecat==0
RR estimate, and lower and upper 95% confidence limits
------------------------------------------------------------
         RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
      2.773       21.05        0.0000         1.760       4.371
------------------------------------------------------------

. 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
------------------------------------------------------------
         RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
      4.578       55.78        0.0000         2.952       7.101
------------------------------------------------------------

. 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
------------------------------------------------------------
         RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
      6.688       93.81        0.0000         4.286      10.438
------------------------------------------------------------

. 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
------------------------------------------------------------
         RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
     17.111      210.78        0.0000        10.114      28.949
------------------------------------------------------------
```

Only the second age group is not significantly different from the first one.

*Q6

To obtain the results shown in the lecture:

```
. strate grade, per(1000)
Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals
(1677 records included in the analysis)
    grade      _D         _Y      _Rate     _Lower     _Upper
        1     221   20.3398     10.865      9.523     12.397
        2     182    7.2656     25.050     21.662     28.967

. stmh grade
Maximum likelihood estimate of the rate ratio
  comparing grade==2 vs. grade==1
RR estimate, and lower and upper 95% confidence limits
------------------------------------------------------------
         RR        chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
      2.305       73.78        0.0000         1.895       2.805
------------------------------------------------------------
```

*Q7 - Q9

To examine whether the estimated RR for Low grade, RR=2.31 (1.90, 2.81), is confounded by age at entry, we examine the stratum specific estimates and assess whether they differ only because of random variation. The test for effect modification is not significant (P=0.79) and the values of the stratum-specific RRs are quite similar (they range from 1.2 to 1.9 with no evidence of a trend). Hence we summarise them using the Mantel-Haenszel estimator to obtain RR=1.43 (1.16, 1.76). The result shows that the crude estimate of the effect of grade was partly confounded by age (at entry).

```
. stmh grade, by(agecat)
Maximum likelihood estimate of the rate ratio
  comparing grade==2 vs. grade==1
  by agecat

RR estimate, and lower and upper 95% confidence limits
   agecat       RR    Lower    Upper
      40-     1.22     0.42     3.57
      45-     1.36     0.67     2.75
      50-     1.92     1.23     3.01
      55-     1.43     1.00     2.06
      60-     1.21     0.82     1.80
      65-     1.40     0.54     3.62

Overall estimate controlling for agecat
        ------------------------------------------------------------
             RR        chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          1.429       11.36        0.0008         1.160       1.761
        ------------------------------------------------------------
Approx test for unequal RRs (effect modification): chi2(5) =     2.44
                                                   Pr>chi2 =   0.7854
```

*Q10

We now examine the effect of Low grade on CHD mortality: First we need to redefine `stset`:

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

Then we compute the RR for grade:

```
. stmh grade
RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR        chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          1.991       18.44        0.0000         1.445       2.743
        ------------------------------------------------------------
```

The estimated (crude) RR is 1.99 (1.45,2.74). Stratifying by smoking we find:

```
. stmh grade, by(smok)
RR estimate, and lower and upper 95% confidence limits
    smok       RR    Lower    Upper
       1     3.88     1.18    12.72
       2     1.64     0.93     2.91
       3     2.03     1.04     3.94
       4     1.33     0.70     2.56
       5     1.93     0.72     5.18
```

```
Overall estimate controlling for smok
------------------------------------------------------------
        RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
     1.761      12.06        0.0005         1.274       2.435
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(4) =     2.76
                                                   Pr>chi2 =   0.5988
```

Although the stratum specific estimate vary from 1.6 to 3.9, the variation is not systematic and the test for effect modification is not significant. The Mantel-Haenszel estimate is RR= 1.76 (1.27, 2.44) indicating that the original estimate of the RR for CHD mortality was partly confounded by smoking.

Key points:

1. To analyse data from cohort studies using Stata, we need first 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 whether a potential confounder "explains away" part , or all, of an estimated RR we do the following:
   - We stratify the data according to categories of the confounder and we 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 we 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 we DO NOT use the Mantel-Haenszel summary estimate but we report the stratum specific RRs.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 3 SOLUTIONS

1.
Open a log file, change directory and read the data:

```
. log using sme3
. cd h:\sme
. use trinmlsh
. describe
. sum
```

The variables **death** and **cvdeath** indicate respectively whether a subject has died during the follow-up and whether he has died of CV causes. The variables **years** and **y** are identical: they both indicate the follow-up time expressed in years. The variable **days** holds instead the same information in days. **timein** and **timeout** are date variables that hold the date of entry and date of exit into/from the study. Their difference is equal to the **days** variable.

2.
The Lifetable estimates, produced at yearly intervals, are:

```
. ltable years death,g noconf saving(plot1,replace) yscale(range(0 1))
title(Lifetable survival estimate)
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	1	318	16	0	0.9497	0.0123	0.9192	0.9689
1	2	302	11	0	0.9151	0.0156	0.8786	0.9410
2	3	291	7	1	0.8930	0.0173	0.8536	0.9224
3	4	283	14	1	0.8488	0.0201	0.8044	0.8838
4	5	268	12	1	0.8107	0.0220	0.7631	0.8497
5	6	255	4	22	0.7974	0.0226	0.7487	0.8377
6	7	229	5	29	0.7788	0.0236	0.7284	0.8211
7	8	195	13	29	0.7227	0.0265	0.6668	0.7709
8	9	153	5	85	0.6900	0.0291	0.6291	0.7431
9	10	63	1	62	0.6685	0.0353	0.5941	0.7323

The plot is shown in the next page.

3.
Before producing the Kaplan-Meier estimates we need to define the time-scale and the outcome of interest:

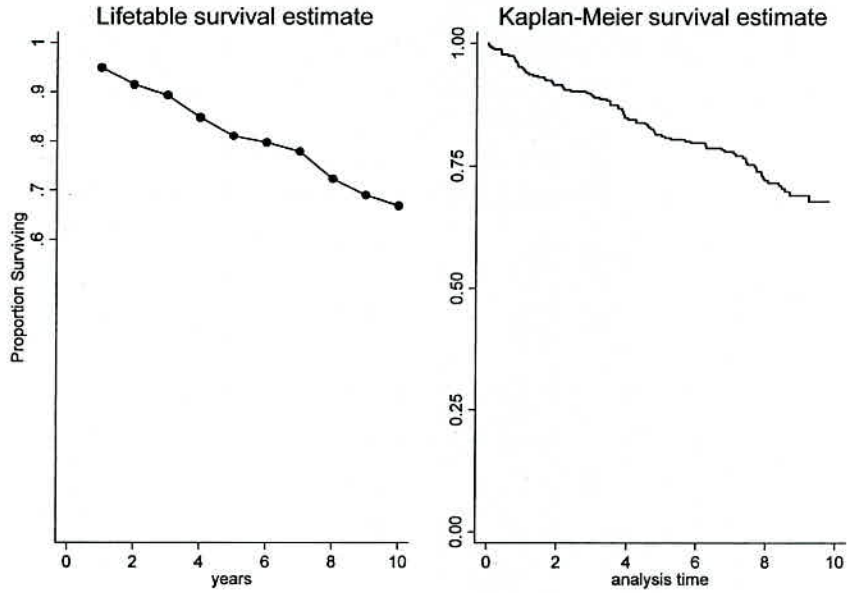
```
stset timeout,f(death) origin(timein) enter(timein) scale(365.25) id(id)

                id:  id
     failure event:  death ~= 0 & death ~= .
obs. time interval:  (timeout[_n-1], timeout]
 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
```

This shows that, out of 318 subject, 88 had died by the end of the follow-up, and that the maximum follow-up time was of about 10 years. The K-M plot is then produced with:

```
. sts gr,saving(plot2,replace)
```

It is shown below, next to that estimated via the Lifetable method.



4.

The graphical comparison shows the following features:

- The two curves are very similar: if you could superimpose them you would see very little difference.
- The Lifetable plot, as produced by Stata does not start from 1.
- With a different dataset, but not this one, some differences could be seen if the intervals used to estimate the Lifetable curve were wider than 1 year (in the Trinidad data the death rate is pretty much constant so that having wider intervals would not give a much worse fit).

Comparing the actual estimated values for the K-M curve (possible using the **`sts list`**

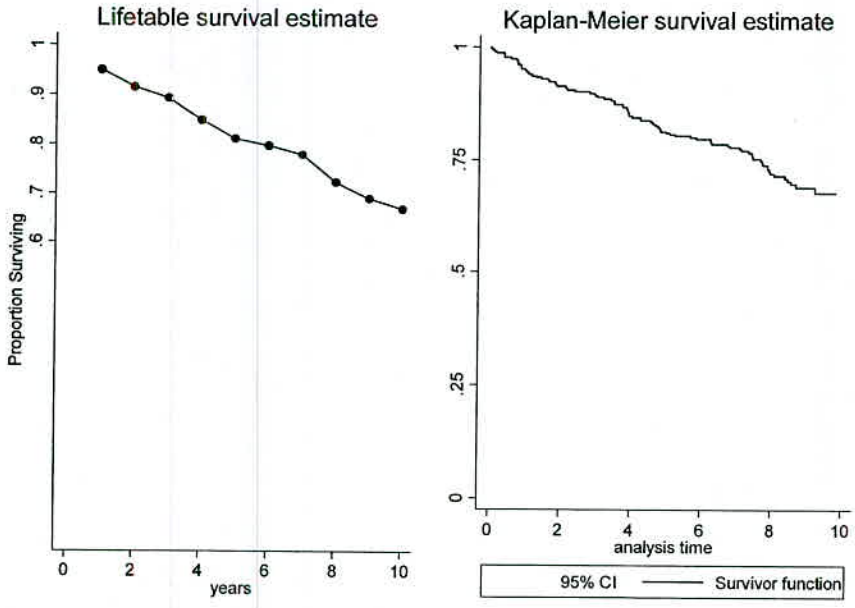
command) with those produced by the Lifetable method we can see how close the values are. For example, an extract of that listing gives:

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1.005	302	1	0	**0.9465**	0.0126	0.9154	0.9664
2.119	291	1	0	**0.9119**	0.0159	0.8750	0.9384
2.971	284	1	0	**0.8930**	0.0173	0.8536	0.9224
3.031	283	1	0	**0.8899**	0.0176	0.8500	0.9197
4.005	268	1	0	**0.8456**	0.0203	0.8010	0.8810
4.977	256	1	0	**0.8107**	0.0220	0.7630	0.8497
5.057	255	0	1	**0.8107**	0.0220	0.7630	0.8497
5.974	230	0	1	**0.7975**	0.0226	0.7488	0.8378

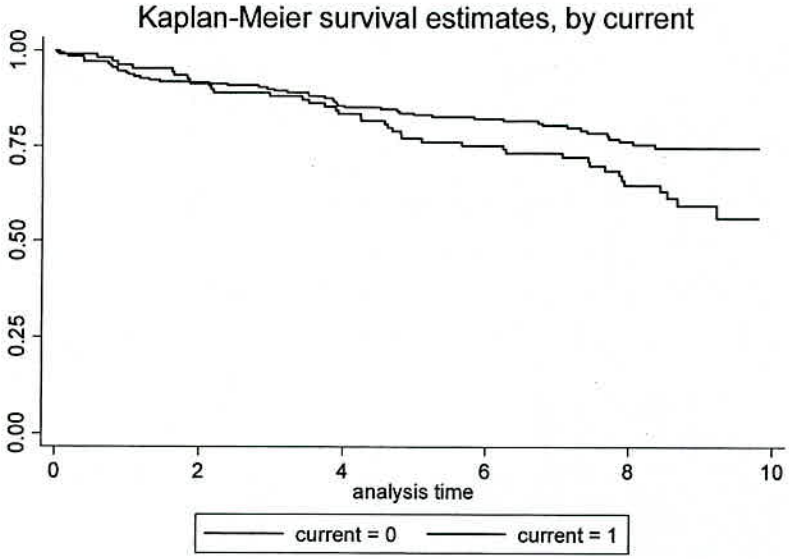
5.

The graphical comparison of the estimated curves with 95% confidence intervals, shown on the next page, highlights the following features:

- Both methods use the Greenwood approximation to compute the confidence intervals.
- For the Lifetable curves the confidence intervals are not joined: this is correct because the calculations are based on just the point estimates, not on the whole curve (strangely, the K-M plot stresses this feature in the title but then joins the points).



6.
Using the K-M method:



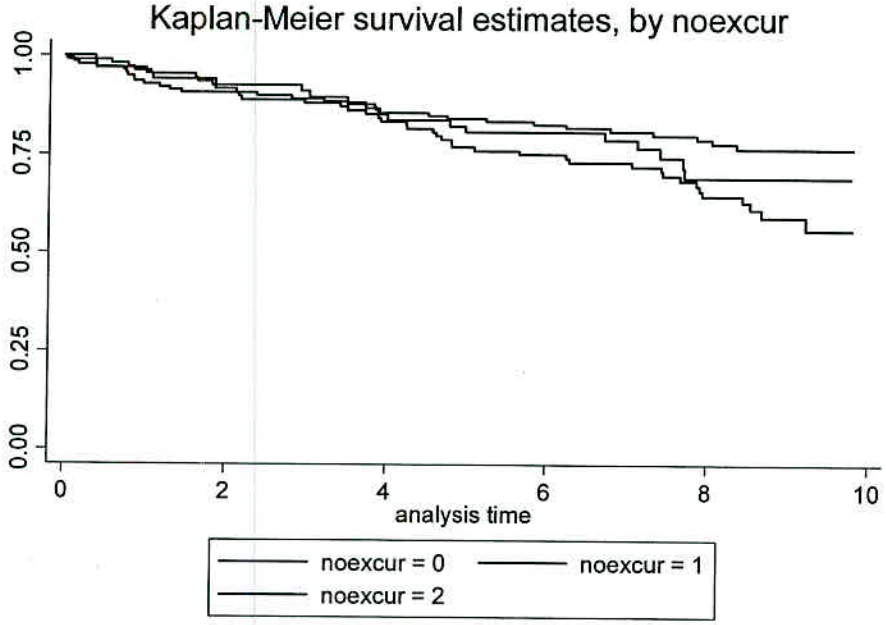
The two curves appear to overlap up to around 4 years and then to separate with the current smokers (at entry into the study) suffering greater mortality.

7.
This is confirmed by the Logrank test which shows that the two curves are significantly different (P=0.03):

```
Log-rank test for equality of survivor functions
------------------------------------------------

         |  Events
current  |  observed       expected
---------+-------------------------
0        |        48          57.61
1        |        40          30.39
---------+-------------------------
Total    |        88          88.00

                chi2(1) =       4.64
                Pr>chi2 =     0.0312
```

8.
Separating the baseline group into non- and ex-smokers shows how the non-smokers differ in mortality from the current smokers, while the ex-smokers are in between, with the Logrank test of their difference being only borderline significant (P=0.06):

```
Log-rank test for equality of survivor functions
----------------------------------------------

        |  Events
noexcur |  observed       expected
--------+-------------------------
0-      |        30          39.54
1-      |        18          18.32
2-      |        40          30.14
--------+-------------------------
Total   |        88          88.00

          chi2(2) =       5.54
          Pr>chi2 =     0.0628
```



Key points:

1. To analyse data from cohort studies while focussing on the time to an event instead of the rate of an event we can use either the Lifetable or the Kaplan-Meier method. They both give a graphical description of the survival pattern of the cohort.

2. These two methods differ only in one aspect: the Lifetable uses data summarised by a given time interval (e.g., yearly) while the Kaplan-Meier uses the individual data on event or censoring times. As a consequence, the first method leads to smooth curves, the second to step-functions.

3. To test whether two survival curves are similar we use the Logrank test. . The result allows to state whether any differences between the two curves are significant. However, it does not allow to quantify the difference.

4. All of these methods are useful for exploring the survival patterns of subjects followed over time and for testing whether such patterns differ. However they do not allow to summarise the data in the same way as rates and RRs do.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 4 SOLUTIONS

Case-Control Studies

1.a) True risk ratio $= \pi_1 \div \pi_0 \quad = 0.01 \div 0.005 \quad = 2.00$

True odds ratio $= \frac{\pi_1}{1-\pi_1} \div \frac{\pi_0}{1-\pi_0} \quad = \frac{0.01}{0.99} \div \frac{0.005}{0.995} \quad = 2.01$

Note that the true risk ratio and true odds ratio are almost identical (because the disease is rare).

b) Rounding off to the nearest control gives

	Exposed	Unexposed	Total
Cases	100	450	550
Controls	55	495	550

$S_H = 550 \div 99450 = 0.0055$ – much smaller than S_D (=1)

$OR = \frac{100 \times 495}{450 \times 55} = 2.00$

The estimated odds ratio is very close to the true risk ratio (in fact it's exactly equal to the true risk ratio if we round off to the nearest control).

c) Rounding off to the nearest control gives

	Exposed	Unexposed	Total
Cases	100	315	415
Controls	41	374	415

$OR = \frac{100 \times 374}{315 \times 41} = 2.90$

We now obtain an overestimate of the odds ratio due to selection bias. (Exposed cases are more likely to be included than unexposed cases so we overestimate the odds of exposure in the cases.)

2. True risk ratio = $0.10 \div 0.05 = 2.00$

True odds ratio $= \frac{0.10}{0.90} \div \frac{0.05}{0.95} = 2.11$

When the disease is rare (question 1) the odds ratio is almost identical to the risk ratio. As the disease becomes less rare, the difference between the odds ratio and the risk ratio increases. The odds ratio is always further away from 1 than the risk ratio.

3. Est. odds ratio $= \frac{131 \times 111}{58 \times 462}$

$= 0.54$

Error factor $= 1.45$

95% c.i. $= (0.37, 0.79)$

$\chi^2 = 10.53$ $p = 0.001$

Women with a current spouse/partner have about half the odds of having HIV infection of women without a current spouse/partner; i.e. women with a current spouse/partner are <u>less</u> likely to be HIV positive. The confidence interval indicates that the true reduction in odds may be as little as 20% or as much as (almost) 70%. It is extremely unlikely that this association is a chance finding.

Key points

- The key feature of case-control studies is that they try to recruit a high proportion of cases but only a small proportion of non-cases.
- We cannot (usually) estimate absolute risks, rates or odds of disease from a case-control study because we only recruit an unknown (usually) fraction of the non-cases.
- Case-control studies estimate the (exposure) odds ratio; when the outcome is rare this is close to the (disease) risk ratio.
- If the probability of recruitment is associated with whether an individual is exposed or not, the odds ratio will be biased.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 5 SOLUTIONS

1. MEASURE OF EFFECT

(i) The diet data are from a cohort study
(ii) The outcome variable is coronary heart disease yes/no (CHD)
(iii) The rate ratio should be used as the measure of effect (person-years of followup are available as we have date of entry to (doe), and exit from (dox), the study).

2. CLASSIFICATION OF VARIABLES

Note that age at entry to the study must be calculated (using dob and doe), but it is implicit that age is an explanatory variable
a) the exposure is energy intake

b)
(i) Height and age are known to be associated with the risk of CHD. They may be associated with energy intake. So height and age are potential confounders when estimating the effect of energy intake on CHD.
(ii) Job may be associated with the risk of CHD, and may be associated with energy intake. Job is a potential confounder.
(iii) Obesity is a known risk factor for CHD, and BMI is a better indicator of obesity than weight as it takes account of height. BMI is likely to be associated with energy intake. So BMI is a potential confounder. (Note that it could be argued that BMI is on a causal pathway linking energy intake to coronary heart disease. However, in this study the hypothesis is that energy intake is a marker of physical activity, and that the risk of CHD will *decrease* with increasing energy intake, rather than that the effect of energy intake is mediated by obesity (which would imply that the risk of CHD will increase with increasing energy intake).)
(iv) month when the dietary record was taken should not be associated with the risk of coronary heart disease (although it may be associated with energy intake). Therefore month is unlikely to be a confounder when estimating the effect of energy intake on coronary heart disease.
(v) Fat intake is a component of energy intake, and is therefore expected to be highly correlated with total energy intake. It is likely to be impossible to separate the 2 effects using the variables as given.

c) It is plausible that the effect of energy intake varies depending on an individual's BMI. So BMI may be considered both as a potential confounder and as a potential effect-modifier.

3. DATA CHECKING AND EDITING

(i) The describe command gives the following output.

```
. desc
Contains data from dietsme2.dta
  obs:           337                          Diet data with dates
 vars:            12                          20 Dec 2001 10:22
 size:        14,154 (99.7% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g                  Subject identity number
doe             long   %dDmCY                 Date of entry
dox             long   %dDmCY                 Date of exit
dob             long   %dDmCY                 Date of birth
fail            int    %8.0g                  Outcome (CHD = 1 3 13)
job             int    %8.0g                  Occupation
```

```
month           byte    %8.0g                 month of survey
energy          float   %9.0g                 Total energy (100 kcals/day)
height          float   %9.0g                 Height (cm)
weight          float   %9.0g                 Weight (kg)
fat             float   %9.0g                 Total fat (10g/day)
chd             byte    %8.0g                 CHD indicator
------------------------------------------------------------------------------
Sorted by:  id
```

There are data on 337 individuals, and there are 12 variables in the data set. All variables are stored as a number. Dates are formatted as day/month/year. The variable labels tell you a bit more about the variable - e.g. energy is "Total energy (100 kcals/day)" while height is measured in cm and weight in kg.

The summarize command gives the following output:

. summ

Variable	Obs	Mean	Std. Dev.	Min	Max
id	337	169	97.42775	1	337
doe	337	1703.237	1247.819	320	13530
dox	337	6548.671	1880.709	-6970	7640
dob	337	-16349.99	2739.338	-23366	-1478
fail	337	2	4.35685	0	15
job	337	1.145401	.8554367	0	2
month	337	6.231454	3.740045	1	12
energy	337	27.93472	5.51852	8	73
height	332	172.3012	9.494772	79	190
weight	333	72.18318	14.2611	7	199
fat	337	12.74542	2.365814	7.26	21.629
chd	337	.1364985	.3438277	0	1

The numbers for the date variables are hard to interpret, as they are measured as days since 1 January 1960.
Job categories range from 0 to 2, months from 1 to 12, energy intake values from 8 to 73 (100 kcals/day), height from 79cm to 190cm, weight from 7kg to 199kg, and fat intake from 7.26 to 21.629 (10g/day). CHD ranges in value from 0 to 1. Mean values are also given - e.g. mean energy intake is 27.9.
332 and 333 individuals respectively have data on height and weight - i.e. 5 and 4 individuals respectively do not have data on height and weight.

(ii) The explanatory variables job and month are categorical variables.
(iii)

. tab job if chd==0

Occupation	Freq.	Percent	Cum.
0	90	30.93	30.93
1	70	24.05	54.98
2	131	45.02	100.00
Total	291	100.00	

. tab job if chd==1

Occupation	Freq.	Percent	Cum.
0	12	26.09	26.09
1	14	30.43	56.52
2	20	43.48	100.00
Total	46	100.00	

The above tables looks sensible - occupation codes are in the allowed categories (0=driver, 1=conductor, 2=bank)

. tab month if chd==0

month of survey	Freq.	Percent	Cum.
1	32	11.00	11.00
2	31	10.65	21.65
3	31	10.65	32.30
4	10	3.44	35.74
5	28	9.62	45.36
6	28	9.62	54.98
7	23	7.90	62.89
8	14	4.81	67.70
9	8	2.75	70.45
10	21	7.22	77.66
11	35	12.03	89.69
12	30	10.31	100.00
Total	291	100.00	

. tab month if chd==1

month of survey	Freq.	Percent	Cum.
1	6	13.04	13.04
2	3	6.52	19.57
3	8	17.39	36.96
4	4	8.70	45.65
5	7	15.22	60.87
6	2	4.35	65.22
7	1	2.17	67.39
8	1	2.17	69.57
9	2	4.35	73.91
10	3	6.52	80.43
11	5	10.87	91.30
12	4	8.70	100.00
Total	46	100.00	

The above makes sense - all months are in the allowed range 1 to 12.

(iv) Energy intake, height, weight, and fat intake are quantitative explanatory variables. Age at entry is a quantitative explanatory variable that needs to be calculated from the variables dob and doe - see part (vi).

(v)

a) energy intake

```
. histogram energy if chd==0, percent start(5) width(5) xlab(5(5)75)
. histogram energy if chd==1, percent start(5) width(5) xlab(5(5)75)
```

```
. list id energy if energy<20 | (energy>=40 & energy!=.)

        id     energy
 8.    123         73
```

```
  42.        84         42
  64.       125         69
  84.        85         42
  95.       230         40
 150.       233          8
 173.       103         18
 174.       170         40
 238.        65         19
 239.       309         18
 265.         2         19
 302.       147         19
 308.       247         18
 309.       249         19
```

The ids particularly worth checking are those with energy intakes appreciably less than 20 or appreciably more than 40 - i.e. ids 123, 125, 233. Checks reveal these to be in error.

```
. replace energy=44 if id==123
. replace energy=39 if id==125
. replace energy=17 if id==233
```

b) height

```
. histogram height if chd==0, percent start(70) width(10) xlab(70(10)190)
. histogram height if chd==1, percent start(70) width(10) xlab(70(10)190)
```

The above histograms show there are 2 individuals with an unusually low height (≤90cm). This is likely to be a data error and should be checked.

```
. list id height if height<=90

            id     height
252.       112         79
277.       323         86
```

Checks reveal the values for ids 112 and 323 to be in error

```
. replace height=179 if id==112
. replace height=187 if id==323
```

c) weight

```
. histogram weight if chd==0, percent start(0) width(10) xlab(0(10)200)
. histogram weight if chd==1, percent start(0) width(10) xlab(0(10)200)
```

The above histograms show there are a few individuals with an unusually low weight (≤40kg), or high weight (>=120kg). These are likely to be data errors and should be checked.

```
. list id weight if weight<=40 | (weight>=120 & weight!=.)

            id     weight
 22.       328        199
 93.       100          7
122.        77         30
163.       330        140
182.       237         18
```

The weights for these 5 ids are in error and should be corrected:

```
. replace weight=100 if id==328
. replace weight=70 if id==100
. replace weight=60 if id==77
. replace weight=98 if id==330
. replace weight=82 if id==237
```

d) fat intake

```
. histogram fat if chd==0, percent start(5) width(5) xlab(5(5)25)
. histogram fat if chd==1, percent start(5) width(5) xlab(5(5)25)
```

It is sensible to check the values of fat intake above 20.

```
. list id fat if fat>20 & fat!=.

            id        fat
 49.       256     20.763
101.       123     21.629
133.        85     20.112
174.       195     20.132
```

All values are actually extremely close to 20, and all are in fact correct.

(vi) Consistency checks

a) dob, doe, and dox, and age at entry

A logical first step is to check whether the dates dob, doe, and dox are consistent - i.e. dob before doe, and doe before dox.

```
. list id dob doe dox if dob>=doe
```

all dates of birth are before date of entry

```
.  list id dob doe dox if doe>=dox

            id        dob        doe        dox
 93.       100  19Jan1934  16Mar1970  01Dec1940
228.       251  15Jun1913  16Feb1992  01Dec1980
279.         8  11Dec1918  16May1969  01Dec1950
283.       303  20Dec1911  16Jan1997  01Dec1980
```

4 individuals have dox before doe - so the entry for one of doe or dox must be in error. These should be checked. For ids 251 and 303, the doe is in error, as there was no recruitment in the 1990s. For ids 8 and 100 the dox is in error, as entry to the study was between 1960 and 1970 (so dox could not be before then).

```
. replace doe=mdy(2,16,1962) if id==251
. replace doe=mdy(1,16,1964) if id==303
. replace dox=mdy(12,1,1980) if id==8
. replace dox=mdy(12,1,1980) if id==100

***to generate age at entry to the study
. gen ageentry=(doe-dob)/365.25
. summ ageentry
. histogram ageentry if chd==0, percent start(0) width(5) xlab(0(5)70)
. histogram ageentry if chd==1, percent start(0) width(5) xlab(0(5)70)
```

These histograms show a few individuals outside the range known to have been recruited to the study (30 to 67 years old). The dates of birth and dates of entry to the study of these individuals should be checked.

```
. list id dob doe ageentry if ageentry<30 | (ageentry>=68 & ageentry!=.)

            id        dob        doe   ageentry
181.       161  15Dec1946  16May1964   17.41821
332.       289  15Dec1955  16Nov1964   8.922656
```

It appears that the dates of birth are probably in error for these 2 individuals (the dates of entry to the study make sense as they are in the 1960s). Checks reveal this to be the case.

```
. replace dob=mdy(12,27,1915) if id==161
. replace dob=mdy(1,14,1911) if id==289

*****recalculate age at entry
. drop ageentry
. gen ageentry=(doe-dob)/365.25
```

b) Other consistency checks

****Consistency between height and weight**

```
. scatter weight height if chd==0
. scatter weight height if chd==1
```

These scatter plots shows 2 individuals for whom height and weight appear to be inconsistent

```
. list id weight height if height>=190 & weight<50

            id     weight     height
154.       139         46        190
```

```
. list id weight height if height<155 & weight>100

            id     weight     height
314.       126        110        152
```

The height of id 139 is in error, and the weight of id 126 is in error.

```
. replace height=160 if id==139
. replace weight=50 if id==126
```

Other checks do not reveal obvious errors. A scatter plot of energy and fat intake confirms their suspected strong association (see q2).

```
. scatter energy weight if chd==0
. scatter energy weight if chd==1
. scatter energy height if chd==0
. scatter energy height if chd==1
. scatter ageentry energy if chd==0
. scatter ageentry energy if chd==1
. scatter fat energy if chd==0
. scatter fat energy if chd==1
. scatter weight fat if chd==0
. scatter weight fat if chd==1
```

(viii) Recheck the data after making corrections

```
. summ energy height weight ageentry

. histogram energy if chd==0, percent start(15) width(5) xlab(15(5)45)
. histogram energy if chd==1, percent start(15) width(5) xlab(15(5)45)

. histogram height if chd==0, percent start(150) width(5) xlab(150(5)200)
. histogram height if chd==1, percent start(150) width(5) xlab(150(5)200)

. histogram weight if chd==0, percent start(40) width(5) xlab(40(5)110)
. histogram weight if chd==1, percent start(40) width(5) xlab(40(5)110)

. histogram ageentry if chd==0, percent start(30) width(5) xlab(30(5)70)
. histogram ageentry if chd==1, percent start(30) width(5) xlab(30(5)70)
```

The distributions look reasonable.

Scatter plots also look reasonable (repeat the commands given above).

(ix) Calculate BMI

```
. gen bmi=weight*100*100/(height*height)
. label variable bmi "BMI"
. histogram bmi if chd==0, percent start(15) width(1) xlab(15(5)35)
. histogram bmi if chd==1, percent start(15) width(1) xlab(15(5)35)
```

The distribution of BMI looks reasonable.

4. DATA REDUCTION

a) energy intake

```
. histogram energy, percent start(15) width(1) xlab(15(5)45)
. centile energy, c(33 67)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
energy	337	33	26	25	26
		67	30	29	30

```
. gen encat=energy
. recode encat min/25=1 26/30=2 31/max=3
. table encat, c(min energy max energy freq)
```

encat	min(energy)	max(energy)	Freq.
1	17	25	107
2	26	30	141
3	31	44	89

b) height

```
. histogram height, percent start(150) width(5) xlab(150(10)200)
. centile height, c(33 67)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]

```
-------------+-------------------------------------------------------------
      height |     332          33         170         170         171
             |                  67         176         175    176.5854
```

```
. gen htcat=height
. recode htcat min/169=1 170/176=2 177/max=3
. table htcat, c(min height max height freq)
```

```
----------------------------------------------
   htcat | min(height)  max(height)       Freq.
---------+------------------------------------
       1 |         152          169         92
       2 |         170          176        147
       3 |         177          190         93
----------------------------------------------
```

c) BMI

```
. histogram bmi, percent start(15) width(1) xlab(15(5)35)
. centile bmi, c(33 67)
```

```
                                                      -- Binom. Interp. --
    Variable |      Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         bmi |      332          33      22.43617        21.83702    22.86252
             |                   67      25.51298         25.1429    26.13816
```

```
. gen bmicat=bmi
. recode bmicat min/23=1 23/26=2 26/max=3
. table bmicat, c(min bmi max bmi freq)
```

```
-------------------------------------------
   bmicat |   min(bmi)     max(bmi)       Freq.
----------+--------------------------------
        1 |   15.88697      22.9854         131
        2 |   23.04002    25.94548         104
        3 |   26.02617    33.49768          97
-------------------------------------------
```

d) age at entry

```
. histogram ageentry, percent start(30) width(5) xlab(30(5)70)
. gen ageentgp=ageentry
. recode ageentgp min/45=1 45/55=2 55/max=3
. table ageentgp, c(min ageentry max ageentry freq)
```

```
-----------------------------------------------------
 ageentgp | min(ageentry)  max(ageentry)         Freq.
----------+------------------------------------------
        1 |      30.07529       44.90349           78
        2 |      45.05955       54.75428          176
        3 |      55.07187       67.09925           83
-----------------------------------------------------
```

****note that age is grouped based on "standard" age bands rather than centiles**

*****Note that these are not the only possible groupings, just ones that are sensible based on histograms of the data and centiles. 3 groups are used for each variable because the number of cases in the data set is relatively small (46), although it is worth considering 4 for the exposure variable, energy.

5A. CROSS-TABULATIONS/RATES OF DISEASE FOR EACH EXPLANATORY VARIABLE

a)
(i)

```
. stset dox, fail(chd) enter(doe) scale(365.25) id(id)

                id:  id
     failure event:  chd ~= 0 & chd ~= .
obs. time interval:  (dox[_n-1], dox]
 enter on or after:  time doe
 exit on or before:  time dox
    t for analysis:  time/365.25

------------------------------------------------------------------------------
      337  total obs.
        0  exclusions
------------------------------------------------------------------------------
      337  obs. remaining, representing
      337  subjects
       46  failures in single failure-per-subject data
 4603.669  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =  .8761123
                                  last observed exit t =  20.91718
```

There were 46 CHD events among the 337 individuals.

(ii)

```
. strate encat, per(1000)

         failure _d:  chd
   analysis time _t:  dox/365.25
  enter on or after:  time doe
  exit on or before:  time dox
                 id:  id
```

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

encat	D	Y	Rate	Lower	Upper
1	23	1.3759	16.7160	11.1082	25.1547
2	16	1.9542	8.1873	5.0158	13.3642
3	7	1.2735	5.4966	2.6204	11.5298

```
. strate htcat, per(1000)

         failure _d:  chd
   analysis time _t:  dox/365.25
  enter on or after:  time doe
  exit on or before:  time dox
                 id:  id
```

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

htcat	D	Y	Rate	Lower	Upper
1	19	1.1532	16.4760	10.5093	25.8304
2	21	1.9238	10.9160	7.1173	16.7422
3	5	1.4566	3.4326	1.4288	8.2470

```
. strate bmicat, per(1000)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
              id:  id

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

   bmicat      D          Y       Rate      Lower      Upper
        1     18     1.7404    10.3426     6.5163    16.4157
        2      9     1.4819     6.0731     3.1599    11.6720
        3     18     1.3113    13.7273     8.6488    21.7879

. strate ageentgp, per(1000)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
              id:  id

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

 ageentgp      D          Y       Rate      Lower      Upper
        1      7     1.2309     5.6871     2.7112    11.9292
        2     25     2.5862     9.6666     6.5318    14.3059
        3     14     0.7866    17.7985    10.5412    30.0523

. strate job, per(1000)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
              id:  id

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

     job      D          Y       Rate      Lower      Upper
       0     12     1.2271     9.7792     5.5537    17.2196
       1     14     1.0435    13.4170     7.9462    22.6542
       2     20     2.3331     8.5722     5.5304    13.2870
```

The risk of CHD appears to decrease with increasing energy intake and with height, and to increase with age at entry. The observed relationship with BMI is U-shaped - lowest for the middle of the 3 categories. Bus conductors had higher rates of CHD than bus drivers or bank workers. We need to calculate measures of effect and to assess whether these patterns are statistically significant.

b)

. tab htcat encat, row chi2

htcat	encat 1	2	3	Total
1	36	35	21	92
	39.13	38.04	22.83	100.00
2	47	61	39	147
	31.97	41.50	26.53	100.00
3	21	44	28	93
	22.58	47.31	30.11	100.00
Total	104	140	88	332
	31.33	42.17	26.51	100.00

Pearson chi2(4) = 5.9747 Pr = 0.201

The trend is for energy intake to increase with height, although a test for heterogeneity is not statistically significant. Height appears to be associated with the risk of CHD (see above).
Height should be considered as a potential confounder when estimating the effect of energy intake on CHD.

. tab bmicat encat, row chi2

bmicat	encat 1	2	3	Total
1	49	54	28	131
	37.40	41.22	21.37	100.00
2	29	48	27	104
	27.88	46.15	25.96	100.00
3	26	38	33	97
	26.80	39.18	34.02	100.00
Total	104	140	88	332
	31.33	42.17	26.51	100.00

Pearson chi2(4) = 6.5770 Pr = 0.160

The trend is for energy intake to increase with BMI, although a test for heterogeneity is not statistically significant.
BMI should be considered as a potential confounder when estimating the effect of energy intake on CHD, although its association with the risk of CHD is not clear (see above).

. tab ageentgp encat, row chi2

ageentgp	encat 1	2	3	Total
1	19	29	30	78
	24.36	37.18	38.46	100.00
2	61	74	41	176
	34.66	42.05	23.30	100.00
3	27	38	18	83
	32.53	45.78	21.69	100.00
Total	107	141	89	337
	31.75	41.84	26.41	100.00

Pearson chi2(4) = 8.1794 Pr = 0.085

Energy intake appears to decline with increasing age. The risk of CHD increases with age in these data (see above). Also, age is known to be a risk factor for CHD. **Age should be controlled for when estimating the effect of energy intake on CHD.**

. tab job encat, row chi2

Occupation	encat 1	2	3	Total
0	34	40	28	102
	33.33	39.22	27.45	100.00
1	26	36	22	84
	30.95	42.86	26.19	100.00
2	47	65	39	151
	31.13	43.05	25.83	100.00
Total	107	141	89	337
	31.75	41.84	26.41	100.00

Pearson chi2(4) = 0.4198 Pr = 0.981

The distribution of energy intake is very similar for each of the 3 jobs.
Job is therefore not expected to confound the association between energy intake and CHD.

5B)

a) energy intake
. stmh encat, c(2,1)

```
        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
               id:  id

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

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.490	5.01	0.0251	0.259	0.927

The rate of CHD in category 2 of energy intake was 0.49 times the rate in category 1, 95% confidence interval [0.26-0.93]

. stmh encat, c(3,1)

```
        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
               id:  id

Maximum likelihood estimate of the rate ratio
  comparing encat==3 vs. encat==1

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.329	7.35	0.0067	0.141	0.766

The rate of CHD in category 3 of energy intake was 0.33 times the rate in category 1, 95% confidence interval [0.14-0.77].

and for a test for linear trend
. stmh encat

```
        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
exit on or before:  time dox
               id:  id
Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.568	8.48	0.0036	0.388	0.831

The rate of CHD is estimated to be reduced by a factor of 0.57 for each 1 unit increase in energy intake, 95% confidence interval [0.39-0.83].

Similarly for the other variables (output not shown):

b) Height

```
. stmh htcat, c(2,1)
. stmh htcat, c(3,1)
. stmh htcat
```

c) BMI

```
. stmh bmicat, c(2,1)
. stmh bmicat, c(3,1)
```

(a test for linear trend is not sensible here, as there is no evidence that the association between BMI and risk of CHD follows a linear trend)

d) Age at entry

```
. stmh ageentgp, c(2,1)
. stmh ageentgp, c(3,1)
```

3) Job

```
. stmh job, c(2,1)
. stmh job, c(3,1)
```

(a test for linear trend is not sensible, as job is not an ordered categorical variable)

Summary

So far, we have learnt that in these data:
a) on crude analysis: energy intake, height, and age appear to be associated with the risk of CHD
b) on crude analysis: BMI and job are not clearly associated with the risk of CHD
c) age should be controlled for when estimating the effect of energy intake on the risk of CHD
d) height should be regarded as a potential confounder when estimating the effect of energy intake on the risk of CHD.
e) although its association with the risk of CHD is not clear in these data, BMI is a known risk factor for CHD. So it should still be considered as a potential confounder of the estimated effect of energy intake on the risk of CHD. It is also still sensible to investigate whether the effect of energy intake on the risk of CHD varies depending on an individual's BMI.
f) Job is unlikely to confound the association between energy intake and the risk of CHD
g) we may be able to model the effect of energy intake on the risk of CHD as a linear trend.

Next steps

Bivariate analyses using stratification with Mantel-Haenszel estimates should be the next step. As energy intake is the only exposure variable, it makes sense to focus on this variable in these analyses.

It is sensible to estimate the effect of energy intake on CHD, controlled for each of the 4 variables age at entry, height, BMI, and job - one at a time.

1) For example, to control for height:

```
. stmh encat if height!=., c(2,1)
. stmh encat if height!=., c(2,1) by(htcat)

. stmh encat if height!=., c(3,1)
. stmh encat if height!=., c(3,1) by(htcat)
```

(output not shown)

and assuming a linear trend for the effect of energy intake on the risk of CHD

```
. stmh encat if height!=.

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.575	7.91	0.0049	0.391	0.846

```
. stmh encat if height!=., by(htcat)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat
  by htcat

RR estimate, and lower and upper 95% confidence limits

    htcat       RR    Lower    Upper
        1     0.62     0.35     1.12
        2     0.63     0.36     1.11
        3     0.57     0.17     1.96

Overall estimate controlling for htcat

------------------------------------------------------------
     RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.624       5.86        0.0155         0.425       0.914
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(2) =     0.02
                                                   Pr>chi2 =   0.9898
```

(note that we need the "if height!=" because we know from earlier descriptive analysis that some individuals do not have data on height. This ensures that the crude estimate of the effect of energy intake, and the estimated effect controlled for height, are based on the same individuals)

The difference between the crude and adjusted (for height) estimate of the effect of energy intake is quite small - so there is little confounding of the estimated effect of energy intake on the risk of CHD by height. However, the estimated effect of energy intake is slightly less strong once we control for height - which is as we would expect, given that height is positively associated with energy intake but negatively associated with the risk of CHD.
Also note that there is no evidence that the effect of energy intake varies depending on an individual's height (the stratum-specific estimates of the effect of energy intake - 0.62, 0.63, and 0.57 - are all extremely similar, and a formal test for interaction is not at all statistically significant (p=0.99)).

2) To control for BMI

```
. stmh encat if bmicat!=., c(2,1)
. stmh encat if bmicat!=., c(2,1) by(bmicat)

. stmh encat if bmicat!=., c(3,1)
. stmh encat if bmicat!=., c(3,1) by(bmicat)
```

(output not shown)

assuming a linear trend for the effect of energy intake on the risk of CHD:

```
. stmh encat if bmicat!=.

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat

RR estimate, and lower and upper 95% confidence limits

------------------------------------------------------------
     RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.575       7.91        0.0049         0.391       0.846
------------------------------------------------------------
```

```
. stmh encat if bmicat!=., by(bmicat)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat
  by bmicat

RR estimate, and lower and upper 95% confidence limits

   bmicat      RR   Lower   Upper
        1    0.46    0.25    0.86
        2    1.06    0.44    2.54
        3    0.48    0.26    0.89

Overall estimate controlling for bmicat

------------------------------------------------------------
     RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.553       8.92        0.0028         0.375       0.816
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(2) =     2.59
                                                   Pr>chi2 =   0.2745
```

There is very little confounding of the estimated effect of energy intake by BMI. The stratum-specific estimates of the effect of energy intake show no evidence of an association between energy intake and the risk of CHD in the intermediate category of BMI, while the effect of energy intake on the risk of CHD is statistically significant in BMI categories 1 and 3. However, a test for interaction indicates that the observed variation in the stratum-specific rate ratios could well have arisen by chance (p=0.27) and the confidence interval for the intermediate category is wide.

3) Controlling for age

```
. stmh encat, c(2,1)
. stmh encat, c(2,1) by(ageentgp)

. stmh encat, c(3,1)
. stmh encat, c(3,1) by(ageentgp)
```

(output not shown)

assuming a linear trend for the effect of energy intake on the risk of CHD:

```
. stmh encat

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat

RR estimate, and lower and upper 95% confidence limits

------------------------------------------------------------
     RR       chi2       P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.568       8.48       0.0036         0.388       0.831
------------------------------------------------------------
```

```
. stmh encat, by(ageentgp)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat
  by ageentgp

RR estimate, and lower and upper 95% confidence limits

 ageentgp      RR   Lower   Upper
        1    0.78    0.30    2.02
        2    0.63    0.37    1.06
        3    0.40    0.19    0.85

Overall estimate controlling for ageentgp
------------------------------------------------------------
     RR       chi2       P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.576       7.68       0.0056         0.390       0.851
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(2) =     1.36
                                                   Pr>chi2 =   0.5069
```

There is little confounding of the effect of energy intake on the risk of CHD by age at entry. There is also no evidence that the effect of energy intake on the risk of CHD varies depending on an individual's age at entry to the study (p=0.51). However, it may be worth noting that the age-specific estimates of the effect of energy intake are greater (rate ratio further away from 1) as age increases.
Note that it would be better to adjust for "current age" rather than age at entry - see ASME.

4) Controlling for job

```
. stmh encat, c(2,1)
. stmh encat, c(2,1) by(job)

. stmh encat, c(3,1)
. stmh encat, c(3,1) by(job)
```

(output not shown)

```
. stmh encat
        failure _d:  chd
  analysis time _t:  dox/365.25
  enter on or after:  time doe
                id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat

RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf.	Interval]
0.568	8.48	0.0036	0.388	0.831

```
. stmh encat, by(job)

        failure _d:  chd
  analysis time _t:  dox/365.25
enter on or after:  time doe
               id:  id

Score test for trend of rates with encat
  with an approximate estimate of the
  rate ratio for a one unit increase in encat
  by job

RR estimate, and lower and upper 95% confidence limits

     job       RR    Lower    Upper
       0     0.51     0.24     1.07
       1     0.58     0.30     1.15
       2     0.60     0.33     1.08

Overall estimate controlling for job

------------------------------------------------------------
     RR       chi2        P>chi2      [95% Conf. Interval]
------------------------------------------------------------
  0.571       8.37        0.0038         0.390       0.835
------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(2) =     0.12
                                                   Pr>chi2 =   0.9431
```

As expected based on univariate analysis, there is little confounding of the estimated effect of energy intake on the risk of CHD by occupation - both the crude and adjusted estimates of the rate ratio are 0.57. There is also no evidence that the effect of energy intake on the risk of CHD varies depending on an individual's occupation (p=0.94).

In summary, there is strong evidence of an association between energy intake and CHD mortality. The association cannot be explained by confounding by occupation, height, BMI, or age at entry to the study - with a caveat that this should be confirmed in a regression model controlling for more than one potential confounder at a time.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 6 SOLUTIONS

1

```
. use mwanza
```

Initalize the new variable and recode it:

```
. generate ed2=ed
. recode ed2 3/4=2
```

Check the recoding worked as wanted:

```
. tabulate ed2 ed
```

ed2	Education 1	2	3	4	Total
1	312	0	0	0	312
2	0	75	365	11	451
Total	312	75	365	11	763

Similarly for age (output not shown):

```
. generate age2=age1
. recode age2 3=2 4/5=3 6=4
. tabulate age2 age1
```

The categories of education and age are now as stated in Q1.

2. To get the simple 2x2 table:

```
. tabulate case ed2, row
```

Case/control	ed2 1	2	Total
0	263	311	574
	45.82	54.18	100.00
1	49	140	189
	25.93	74.07	100.00
Total	312	451	763
	40.89	59.11	100.00

Row percentages are used because column percentages are affected by the different probabilities of selection for cases and controls.

Compared to the table in the lecture notes, this table has the row order swapped round, and the column order also swapped round. STATA automatically puts the category with a lower code value first. This illustrates the importance of being clear what you are treating as exposure and

which category is a case. Software or textbooks may have conventions about row and column ordering, but your data may not come out that way if your codings are different!

To calculate odds ratios:

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

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

------------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   0.413878      23.25        0.0000         0.285782   0.599391
------------------------------------------------------------------------

. mhodds case ed2, c(2,1)

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

------------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   2.416169      23.25        0.0000         1.668360   3.499168
------------------------------------------------------------------------
```

The first of the above two commands considers level 2 (ie higher level of education) as the baseline. The second command is the other way round, so the odds ratio is the reciprocal (one over) the first odds ratio. The second odds ratio is the one in the notes.

To get the χ^2:

```
. tab case ed2, chi exact
Case/control ed2
                    1          2 |     Total
-----------+----------------------+----------
         0 |       263        311 |       574
         1 |        49        140 |       189
-----------+----------------------+----------
     Total |       312        451 |       763

          Pearson chi2(1) =  23.2789   Pr = 0.000
           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000
```

In fact, there's no need to do the exact test since the smallest expected value is 312x189/763=77.3, much bigger 5.

Both exact and approximate tests indicate strong evidence against the null hypothesis.

To obtain the stratified tables we can use `by`, although first we need to `sort`:

```
. sort age2
. by age2: tabulate case ed2, row

-> age2=        1
Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |        18         78 |        96
           |     18.75      81.25 |    100.00
-----------+----------------------+----------
         1 |         4          9 |        13
           |     30.77      69.23 |    100.00
-----------+----------------------+----------
     Total |        22         87 |       109
           |     20.18      79.82 |    100.00

-> age2=        2
Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |        48        144 |       192
           |     25.00      75.00 |    100.00
-----------+----------------------+----------
         1 |        14         82 |        96
           |     14.58      85.42 |    100.00
-----------+----------------------+----------
     Total |        62        226 |       288
           |     21.53      78.47 |    100.00

-> age2=        3
Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |       115         77 |       192
           |     59.90      40.10 |    100.00
-----------+----------------------+----------
         1 |        19         44 |        63
           |     30.16      69.84 |    100.00
-----------+----------------------+----------
     Total |       134        121 |       255
           |     52.55      47.45 |    100.00

-> age2=        4
Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |        82         12 |        94
           |     87.23      12.77 |    100.00
-----------+----------------------+----------
         1 |        12          5 |        17
           |     70.59      29.41 |    100.00
-----------+----------------------+----------
     Total |        94         17 |       111
           |     84.68      15.32 |    100.00
```

To get the odds ratios for the above tables, use `mhodds` with the `by` option

```
. mhodds case ed2, by(age2)

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

---------+-------------------------------------------------------------------
    age2 | Odds ratio        chi2(1)         P>chi2       [95% Conf. Interval]
---------+-------------------------------------------------------------------
       1 |   0.519231           1.02         0.3132         0.142000  1.898596
       2 |   1.952381           4.10         0.0430         1.009050  3.777605
       3 |   3.458647          16.76         0.0000         1.836539  6.513465
       4 |   2.847222           3.05         0.0808         0.833046  9.731360
---------+-------------------------------------------------------------------

   Mantel-Haenszel estimate controlling for age2
   ------------------------------------------------------------
    Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
   ------------------------------------------------------------
      2.289748      17.94        0.0000         1.543506   3.396777
   ------------------------------------------------------------

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

The chi-squared 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 a suspicion 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 values retained. 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.

4. To remind ourselves how religion, and case status, is coded, type:

```
help mwanza
```

(Output not shown.) Religion (`rel`) has one missing value (code 9). To set it to missing:

```
recode rel 9=.
```

To do an exploratory tabulation (remember, cases are code 1, controls are code 0):

```
. tabulate case rel, chi row

Case/contr |                  Religion
        ol |         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
```

Religion is associated with HIV infection. 'Other' religions (code 4) are clearly under-represented among cases relative to controls. Note: row % are used,not column % because cases and controls had different chances of selection.

One could also calculate odds ratios for having HIV infection comparing each religious group in turn with (eg) Moslems – the resulting ORs are 0.57, 0.51, 0.81 respectively.

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

```
. tabulate ed2 rel, chi row

           |                  Religion
       ed2 |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |        19         98         54        141 |       312
           |      6.09      31.41      17.31      45.19 |    100.00
-----------+--------------------------------------------+----------
         2 |        29        223        151         47 |       450
           |      6.44      49.56      33.56      10.44 |    100.00
-----------+--------------------------------------------+----------
     Total |        48        321        205        188 |       762
           |      6.30      42.13      26.90      24.67 |    100.00

          Pearson chi2(3) = 122.6887   Pr = 0.000
```

Religion is also associated with education, 'other' religions being least likely to have education, and Protestants most.

To do the analysis suggested in the practical:

```
. mhodds case ed2 if rel!=., by(rel) c(2,1)
```

The option `c(2,1)` is not strictly necessary here, as the default option in STATA is to compare the higher-numbered code category with the lower one, but is included as a reminder

to think through which category you want to be the reference group before entering the command. The output is as follows:

```
. mhodds case ed2 if rel!=., by(rel) c(2,1)

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

---------+-------------------------------------------------------------------
     rel | Odds ratio       chi2(1)        P>chi2       [95% Conf. Interval]
---------+-------------------------------------------------------------------
       1 |   2.022222          1.29        0.2562         0.584714   6.993816
       2 |   2.252252          7.69        0.0056         1.248565   4.062776
       3 |   1.393519          0.79        0.3745         0.667751   2.908110
       4 |   2.019724          2.15        0.1425         0.774144   5.269413
---------+-------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for rel
    ------------------------------------------------------------
     Odds ratio    chi2(1)        P>chi2       [95% Conf. Interval]
    ------------------------------------------------------------
       1.914248      10.89        0.0010         1.292931   2.834138
    ------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(3)    =     1.03
                                     Pr>chi2    =   0.7931
```

(If you like, try the same command but without the option if `rel!=.`. This will include the missing values as a stratum. In this example, since there is only one record with missing religion, it is not possible to calculate an odds ratio for the 'missing' stratum, so the stratified results are the same either way (STATA prints a message to this effect). However, if there are many missing values, the 'missing' stratum would contribute, and affect the analysis if they were not excluded (see Q6).)

The stratum-specific ratios look fairly similar (2.02, 2.25, 1.29,2.02). The third group seems to have a lower odds ratio than the others but the 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 the group at lowest risk of infection were also the group 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. In this example, we should find the crude odds ratio omitting the person whose religion is unknown:

```
. mhodds case ed2 if rel!=., c(2,1)

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

------------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   2.423963      23.42        0.0000         1.673565   3.510826
------------------------------------------------------------------------
```

When there is only one missing case it won't make much difference in a large sample, but it can make a substantial difference where there are several missing values, and it is good practice always to check that the crude and adjusted estimates apply to the same population.

5. To do the test for trend we type:

```
. tabodds case ed

-----------+------------------------------------------------------------
        ed |      cases     controls       odds       [95% Conf. Interval]
-----------+------------------------------------------------------------
         1 |         49          263    0.18631        0.13734   0.25275
         2 |         24           51    0.47059        0.28969   0.76444
         3 |        110          255    0.43137        0.34495   0.53945
         4 |          6            5    1.20000        0.36623   3.93196
-----------+------------------------------------------------------------
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
```

We are shown two χ^2 tests here. The first one, on three degrees of freedom, tells us that there is evidence for some kind of association between HIV and education. The second one, on one degree of freedom, tells us there is evidence for a more specific kind of association: that the odds of HIV increase by a constant factor for each category of `ed`.

The χ^2 statistic for departure from trend is the difference between the two ones in the output: 26.70-22.24=4.46, on 3-1=2 degrees of freedom. We can look up the p value with the following commands:

```
scalar deptrend=chiprob(2, 4.46)
scalar list deptrend

deptrend =  .10752843
```

So there is no evidence of departure of the log-odds from a linear increase per unit level of `ed`. In other words, a constant multiplicative increase, for each level of `ed`, satisfactorily explains the pattern in odds of HIV.

Remember that this is a case control study so that the ratio of cases to controls (D/H) 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.

To do the test another way we can 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)
```

Odds ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
1.504866	22.24	0.0000	1.269793	1.783457

Note that the χ^2 value is the same as for the score test above.

The estimate of the odds ratio tells us that for each step-up in education category (ed) the odds of having HIV infection is multiplied by 1.5. To be more specific, the odds of having HIV infection for women in Mwanza with 1-3 years education is 1.5 that of women with no education; the odds of having HIV infection for those who have 4-6 years of education is 1.5 time that of those with 1-3 years of education.

6 If those with unknown numbers of sexual partners are included, the crude and adjusted ORs for ed2 are 2.42 and 2.52 respectively. Note that, as before, the missing values form their own stratum in the calculation. However, unlike Q4, there are enough values for this stratum to have its own OR and hence contribute to the MHOR.

If missing values in npa are excluded, the crude and adjusted values are 2.31 and 2.42 respectively.

Both these pairs of results suggest slight confounding, but if you compared the crude result which includes missing values (2.42) and the adjusted one which excludes them (2.42) you would conclude there was no confounding at all.

It is important to be clear who your analysis covers and that comparisons are made within the same group.

7 We should exclude missing values in npa throughout. Assuming they are still coded '.' as in Q6,we can simply type:

```
. tabodds case npa

-----------+------------------------------------------------------------
       npa |      cases     controls       odds      [95% Conf. Interval]
-----------+------------------------------------------------------------
         1 |         27          173    0.15607        0.10404   0.23413
         2 |         92          277    0.33213        0.26235   0.42047
         3 |         40           83    0.48193        0.33048   0.70278
         4 |         24           19    1.26316        0.69194   2.30592
-----------+------------------------------------------------------------
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

. 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)

------------------------------------------------------------------------
 Odds ratio     chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   1.885332       37.26        0.0000         1.538023   2.311068
------------------------------------------------------------------------
```

Note that missing values do not form a separate stratum when npa is the exposure variable in tabodds, unlike when it is a possible confounder in mhodds (as in the previous question). We can see this because a) there is not a separate '.' stratum in the table, and b) the χ^2 has only three degrees of freedom, not four. If we want to treble-check we can repeat the commands with `if npa!=.`

The odds increases by 89% (a factor of 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. As before, the departure from the log-linear trend is small.

KEY POINTS

- Obtain a crude estimate first and then see how it compares with adjusted ones to decide whether there is confounding
- You need to know your data, check any recodings, and 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.
- The STATA command tabodds can be used to look at trends but does not give exact odds in case control studies.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 8 SOLUTIONS

Likelihoods

1. Use the **samex** option to keep scales the same. More data ⇒ more precise estimate.

2. A cut point of 0.2585 corresponds to a 90% c.i. (instead of 90%). So the c.i. is narrower (but we are less sure about it).

3. With less extreme splits the likelihood ratio is roughly symmetric and looks similar to the Gaussian (normal) likelihood ratio. As the split becomes extreme, with only a small number of cases, the likelihood ratio ceases to resemble the Gaussian likelihood.

4. When there are no cases the likelihood (ratio) has no turning point.

5. Again, more data ⇒ more precise estimate.

6. As with the binomial likelihood, when the number of events/cases is small the likelihood (ratio) is not at all symmetric.

7. Most likely value of $\pi = 15/25 = 0.6$. The data are compatible with the hypothesis that the true value of π is 0.5.

```
. blik 15 10, null(0.5) pval
Most likely value for pi  0.60000
Null value for pi         0.50000
Lik ratio for null value          0.60448
Approx pvalue   0.3157
```

8. The most likely value remains 0.6 (24/40). Although the p-value has become smaller (most likely value has changed but the sample is larger), the data are still compatible with the hypothesis that the true value of π is 0.5.

```
. blik 24 16, null(0.5) pval
Most likely value for pi  0.60000
Null value for pi         0.50000
Lik ratio for null value          0.44690
Approx pvalue   0.2044
```

9. The most likely value remains 0.6 (36/60). More data have decreased the p-value further, but data are still compatible with the hypothesis that the true value of π is 0.5.

```
. blik 36 24, null(0.5) pval
Most likely value for pi  0.60000
Null value for pi         0.50000
Lik ratio for null value          0.29876
Approx pvalue   0.1201
```

10.

```
. bloglik 4 6
Most likely value for param     0.40000
cut-point -1.921
Likelihood based limits for param  0.14562  0.70004
Approx quadratic limits for param  0.09634  0.70366
```

Output presents the maximum likelihood estimate ("most likely value") for π (= 4/10 = 0.4). It also presents the 95% c.i. (cut-point = -1.921) obtained from the true log likelihood ratio curve (in yellow on the graph) and from the quadratic approximation to the log likelihood ratio (in red on the graph). Even this very small sample (of 10) the approximation is not too bad.

11.

```
. bloglik 40 60

Likelihood based limits for param  0.30741  0.49766
Approx quadratic limits for param  0.30398  0.49602
```

With the larger sample size there is now very good agreement between the approximate confidence interval the the "exact" c.i. Notice also that with the increased sample size the c.i. has become narrower.

12.

```
. bloglik 400 600
Likelihood based limits for param  0.36992  0.43059
Approx quadratic limits for param  0.36963  0.43037
```

With 400 failures and 600 survivors the approximation is almost perfect (see also graph).

13.

```
. bloglik 2 18
Most likely value for param     0.10000
cut-point -1.921
Likelihood based limits for param  0.01736  0.27796
Approx quadratic limits for param  -0.03149  0.23149
```

The approximation is now quite poor because the true log likelihood ratio curve is far from quadratic in shape. The lower limit of the approximate c.i. is negative, which is not a possible value. With 20 and 180 the approximation is much better and the problem of the negative lower limit has disappeared.

14.

```
 bloglik 2 18,log
Most likely value for param     -2.19722
cut-point -1.921
Likelihood based limits for param  -4.03622  -0.95459
Approx quadratic limits for param  -3.65820  -0.73625

Back on original scale
Most likely value for param     0.10000
Likelihood based limits for param  0.01736  0.27796
Approx quadratic limits for param  0.02513  0.32382
```

By working with the log(odds) parameter and then converting back to π at the end we avoid the problem of a negative lower limit to the approximate confidence interval.

15.

```
. ploglik 7 500

ALL RATES PER    1000
Most likely value for rate parameter      14.00
cut-point -1.921
Likelihood based limits for rate parameter     6.02    27.08
Approx quadratic limits for rate parameter     3.63    24.37
```

16.

```
. ploglik 7 500,log

ALL RATES PER    1000
Most likely value for log rate parameter       2.64
cut-point -1.921
Likelihood based limits for log rate parameter     1.79     3.30
Approx quadratic limits for log rate parameter     1.90     3.38

Back on original scale
Most likely value for rate parameter      14.00
Likelihood based limits for rate parameter     6.02    27.08
Approx quadratic limits for rate parameter     6.67    29.37
```

Working with the log(rate) parameter, our approximate c.i. is closer to the "exact" c.i." than when with worked with the rate parameter.

17.

```
. ploglik 1 1000

ALL RATES PER    1000
Most likely value for rate parameter       1.00
cut-point -1.921
Likelihood based limits for rate parameter     0.06     4.40
Approx quadratic limits for rate parameter    -0.96     2.96

. ploglik 1 1000,log


ALL RATES PER    1000
Most likely value for log rate parameter       0.00
cut-point -1.921
Likelihood based limits for log rate parameter    -2.86     1.48
Approx quadratic limits for log rate parameter    -1.96     1.96

Back on original scale
Most likely value for rate parameter       1.00
Likelihood based limits for rate parameter     0.06     4.40
Approx quadratic limits for rate parameter     0.14     7.10
```

Because the number of events is small (1) the approximations are not very good. Working with the log(rate) avoids the problem of a negative lower limit to the approximate confidence interval.

Key points

- The more data we have, the narrower the likelihood (ratio) curve becomes and hence the more precise our estimates become (narrower confidence intervals).
- The more data we have the more symmetric the likelihood (ratio) curve becomes (and hence the more it resembles the Gaussian likelihood (ratio)).
- The quadratic approximation to the log likelihood ratio for the binomial and Poisson models gets better the more data we have. Even with relatively few data it can be quite good.
- Working with the log(odds) (rather than the risk) and the log(rate) (rather than the rate) tends to improve the approximation and avoids the problem of parameter values which are not meaningful.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 9 SOLUTIONS

For most questions, the output is explained in the notes for the practical.

When tabulating variables for a cross-sectional study, use column percentages if the explanatory variable is the column variable, row percentages if it is the row variable. For example:

tab mf agegrp, col

Microfil. infection	Age group 0	1	2	3	Total
0	156	119	125	80	480
	77.23	54.59	29.48	17.47	36.87
1	46	99	299	378	822
	22.77	45.41	70.52	82.53	63.13
Total	202	218	424	458	1302
	100.00	100.00	100.00	100.00	100.00

13. To examine the association between microfilarial infection and sex:

tab mf sex, col

Microfil. infection	Sex 0	1	Total
0	190	290	480
	30.84	42.27	36.87
1	426	396	822
	69.16	57.73	63.13
Total	616	686	1302
	100.00	100.00	100.00

Column percentages show that infection is more common among males (sex=0) than females (sex=1).

To use logistic regression to examine the association between microfilarial infection and sex:

a) `. xi: logistic mf i.sex`

```
i.sex                 _Isex_0-1     (naturally coded; Isex_0 omitted)

Logit Estimates                                   Number of obs =   1302
                                                  LR chi2(1)    =  18.32
                                                  Prob > chi2   = 0.0000
Log Likelihood = -847.87095                       Pseudo R2     = 0.0107

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Isex_1 |   .6090335   .0709832    -4.255   0.000     .4846555    .7653309
------------------------------------------------------------------------------
```

b) `. est store C`

c) `. logistic mf`

```
Logit Estimates                                   Number of obs =   1302
                                                  LR chi2(0)    =   0.00
                                                  Prob > chi2   =      .
Log Likelihood = -857.02925                       Pseudo R2     = 0.0000

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------
```

d) `. est store D`

e) `. lrtest C D`

```
likelihood-ratio test                             LR chi2(1)  =     18.32
(Assumption: D nested in C)                       Prob > chi2 =    0.0000
```

There is a highly statistically significant association between **sex** and the odds of microfilarial infection. Since **sex** is coded as 0 (males) and 1 (females), it makes no difference whether it is specified as a categorical variable.

14. Logistic regression model with both **area** and **agegrp** as explanatory variables:

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

```
i.agegrp              _Iagegr_0-3   (naturally coded; Iagegr_0 omitted)

Logit Estimates                                   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	8.181	0.000	2.354217	4.037975
_Iagegr_1	2.599132	.5771594	4.301	0.000	1.681945	4.016473
_Iagegr_2	9.76541	2.033437	10.944	0.000	6.49301	14.68706
_Iagegr_3	17.64158	3.808709	13.295	0.000	11.55496	26.93437

The odds ratio for the effect of area has **increased** from 2.41 to 3.08. The interpretation of this odds ratio will be discussed in session 11.

Summary of main 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 microfilarial infection increased with age from 23% in those aged 5-9 to 83% in those aged ≥40.

2. Logistic regression models the log odds of disease and produces odds ratios. We 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** `agegroup`.

3. The **Wald test** is based on the log OR divided by its SE. It tests the null hypothesis that the true odds ratio is 1. The Wald p-value may be obtained directly from the logistic regression output. The **LRT** 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 3 Wald p-values (one for each OR).

Summary of STATA commands

1. Use the **logit** command to obtain the baseline 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. We can get the baseline odds of disease by putting in a term instead of the constant. This involves generating a constant term (`gen const=1`), and then using the `logit` command to perform the logistic regression with this variable, but suppressing the constant in the model (`logistic mf area const , noconstant or`)

3. Indicator variables are used to obtain odds ratios for variables with more than two levels. Stata does this using **xi:** and **i.** as follows: **xi: logistic mf i.agegrp**

4. The LRT is obtained using 5 commands:

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

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 10: Solutions

1.

```
tab area sex,row chi
```

```
   Area of |          Sex
 residence |         0          1 |     Total
-----------+----------------------+----------
         0 |       247        301 |       548
           |     45.07      54.93 |    100.00
-----------+----------------------+----------
         1 |       369        385 |       754
           |     48.94      51.06 |    100.00
-----------+----------------------+----------
     Total |       616        686 |     1,302
           |     47.31      52.69 |    100.00

          Pearson chi2(1) =   1.9027   Pr = 0.168
```

There is a slightly higher proportion of females (`sex=1`) in the savanna area (`area=0`) than in the forest area (55% vs 51%), but this difference is compatible with sampling variation.

2.

```
sort area
by area: tab mf sex,col chi
```

```
-> area = 0

Microfil. |          Sex
 infection |         0          1 |     Total
-----------+----------------------+----------
         0 |        97        170 |       267
           |     39.27      56.48 |     48.72
-----------+----------------------+----------
         1 |       150        131 |       281
           |     60.73      43.52 |     51.28
-----------+----------------------+----------
     Total |       247        301 |       548
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  16.0785   Pr = 0.000
```

```
-> area = 1

Microfil. |          Sex
 infection |         0          1 |     Total
-----------+----------------------+----------
         0 |        93        120 |       213
           |     25.20      31.17 |     28.25
-----------+----------------------+----------
         1 |       276        265 |       541
           |     74.80      68.83 |     71.75
-----------+----------------------+----------
     Total |       369        385 |       754
           |    100.00     100.00 |    100.00
```

```
Pearson chi2(1) =   3.3082   Pr = 0.069
```

In both areas there is some evidence that mf infection is more common in males than females. So sex appears to be an independent (of area) risk factor for mf infection, but does not appear to be strongly associated with area. Hence not likely to be a strong confounder (associated with outcome but not (strongly) with disease.

3.

```
. mhodds mf area
Maximum likelihood estimate of the odds ratio
Comparing area==1 vs. area==0
------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
   2.413363      57.11        0.0000         1.906097   3.055626
------------------------------------------------------------------

. mhodds mf area,by(sex)
Maximum likelihood estimate of the odds ratio
Comparing area==1 vs. area==0
by sex
-------------------------------------------------------------------------------
      sex | Odds Ratio      chi2(1)        P>chi2       [95% Conf. Interval]
----------+--------------------------------------------------------------------
        0 |   1.919140        13.71        0.0002         1.35069    2.72682
        1 |   2.865776        44.28        0.0000         2.07137    3.96486
-------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for sex
    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       2.395256      55.44        0.0000         1.889241   3.036802
    ------------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(1)    =    2.71
                                     Pr>chi2    =  0.0997
```

The stratum-specific ORs look a bit different but the test for interaction tells us that the observed difference is reasonably compatible with sampling variation (p=0.10). I.e. no strong evidence that sex is an effect modifier and it is reasonable to calculate a summary estimate. The sex-adjusted odds ratio (2.40) is very close to the crude OR (2.41) indicating that sex is not an important confounder.

4.

```
. xi:logistic mf i.area i.sex
Logistic regression                               Number of obs   =       1302
                                                  LR chi2(2)      =      73.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -820.18542                       Pseudo R2       =     0.0430

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.396743   .2839052     7.38   0.000     1.900171    3.023083
     _Isex_1 |   .6165796   .0734326    -4.06   0.000     .4882181    .7786899
------------------------------------------------------------------------------
```

The adjusted odds ratio for area (2.40) is almost identical to that obtained from the M-H analysis. There is strong evidence from the Wald test (z=-4.06; p<0.001) that sex is associated with mf infection (i.e. a risk factor) after taking account of area.

```
xi:logistic mf i.area i.sex i.agegrp
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   .425981     8.10   0.000     2.342034    4.032467
     _Isex_1 |   .5591696   .0758462   -4.29   0.000     .4286332    .7294595
  _Iagegrp_1 |   2.567357   .5748449    4.21   0.000      1.65538    3.981758
  _Iagegrp_2 |   10.46237   2.205354   11.14   0.000     6.921597    15.81446
  _Iagegrp_3 |   17.65935   3.834073   13.22   0.000     11.53899    27.02598
------------------------------------------------------------------------------
```

5. The odds ratio associated with **_Iarea_1** (3.07) is the odds ratio for area 1 (forest) vs area 0 (savanna) controlled for any confounding effects of sex and age group. The odds ratio associated with **_Isex_1** (0.56) is the odds ratio for sex 1 (female) vs sex 0 (male) after adjusting for any confounding effects of area and age group. The odds ratio associated with **_Iagegrp_2** (10.46) is the odds ratio for age group 2 (20-39) versus age group 0 (5-9), controlled for any confounding effects of area and sex. Controlling age has changed the estimate of the odds ratio for area (from 2.40 in the previous model, to 3.07), so there is indication that age is a confounder for area. Adding age group to the model has changed the odds ratio for sex to a smaller extent, from 0.62 to 0.56, suggesting only minor confounding of sex by age.

6.

```
xi:logistic mf i.area i.sex i.agegrp
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   .425981     8.10   0.000     2.342034    4.032467
     _Isex_1 |   .5591696   .0758462   -4.29   0.000     .4286332    .7294595
  _Iagegrp_1 |   2.567357   .5748449    4.21   0.000      1.65538    3.981758
  _Iagegrp_2 |   10.46237   2.205354   11.14   0.000     6.921597    15.81446
  _Iagegrp_3 |   17.65935   3.834073   13.22   0.000     11.53899    27.02598
------------------------------------------------------------------------------

estimates store A

xi:logistic mf i.sex i.agegrp
Logistic regression                                     Number of obs   =       1302
                                                        LR chi2(4)      =     278.59
                                                        Prob > chi2     =     0.0000
Log likelihood =  -717.7326                             Pseudo R2       =     0.1625
```

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Isex_1 |   .5559806   .0733739    -4.45   0.000     .4292644    .7201026
  _Iagegrp_1 |   2.818539   .6142679     4.75   0.000     1.838715    4.320495
  _Iagegrp_2 |   8.774755   1.773097    10.75   0.000     5.905197    13.03874
  _Iagegrp_3 |   16.23326   3.409731    13.27   0.000      10.7551    24.50176
------------------------------------------------------------------------------
```

```
. estimates store B

. lrtest B A
likelihood-ratio test                                 LR chi2(1)  =     69.36
(Assumption: B nested in A)                           Prob > chi2 =    0.0000
```

The LRT provides strong evidence against the null hypotheses that, after controlling for age and sex, there is no association between area and mf infection. A similar inference could have been made by examining the Wald test result; $z = 8.10$, $p<0.001$.

7.

```
. xi:logistic mf i.area i.sex
Logistic regression                               Number of obs   =       1302
                                                  LR chi2(2)      =      73.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -820.18542                       Pseudo R2       =     0.0430

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.396743   .2839052     7.38   0.000     1.900171    3.023083
     _Isex_1 |   .6165796   .0734326    -4.06   0.000     .4882181    .7786899
------------------------------------------------------------------------------

estimates store C
lrtest C A
likelihood-ratio test                                 LR chi2(3)  =    274.27
(Assumption: C nested in A)                           Prob > chi2 =    0.0000
```

The LRT indicates that after controlling for area and sex there is strong evidence of an association between age group and mf infection.

Optional note: We cannot use the Wald test statistics provided with the logistic regression output to perform a single Wald test of the null hypothesis that age is not associated with mf infection. However, this is possible to do in Stata.

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

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   3.073138    .425981     8.10   0.000     2.342034    4.032467
     _Isex_1 |   .5591696   .0758462    -4.29   0.000     .4286332    .7294595
  _Iagegrp_1 |   2.567357   .5748449     4.21   0.000      1.65538    3.981758
  _Iagegrp_2 |   10.46237   2.205354    11.14   0.000     6.921597    15.81446
  _Iagegrp_3 |   17.65935   3.834073    13.22   0.000     11.53899    27.02598
------------------------------------------------------------------------------
```

```
. testparm _Iagegrp*
```

```
( 1)  _Iagegrp_1 = 0
( 2)  _Iagegrp_2 = 0
( 3)  _Iagegrp_3 = 0

          chi2(  3) =  223.19
        Prob > chi2 =    0.0000
```

Note that this peforms a chi-squared test based on 3 degrees of freedom, because it is testing all three age group parameters simultaneously (just like the LRT). Both tests lead to very small p-values, indicating strong evidence against the null hypothesis.

8.

Based on the preceding analyses one can conclude that area age and sex are all independently associated with mf infection

9.

```
gen ed2=ed
recode ed2 2/4=2
(ed2: 376 changes made)

tab case ed2,row chi

Case/contr |          ed2
        ol |         1          2 |     Total
-----------+----------------------+----------
         0 |       263        311 |       574
           |     45.82      54.18 |    100.00
-----------+----------------------+----------
         1 |        49        140 |       189
           |     25.93      74.07 |    100.00
-----------+----------------------+----------
     Total |       312        451 |       763
           |     40.89      59.11 |    100.00

          Pearson chi2(1) =  23.2789   Pr = 0.000

. mhodds case ed2
Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs. ed2==1
------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   2.416169      23.25        0.0000         1.668360   3.499168
------------------------------------------------------------------------
```

The crude odds ratio for **ed2** is 2.42 (as observed previously).

10.

```
recode rel 9=.
(rel: 1 changes made)
```

11.

```
mhodds case ed2,by(rel)
Maximum likelihood estimate of the odds ratio
Comparing ed2==2 vs. ed2==1
by rel

note: only 4 of the 5 strata formed in this analysis contribute
      information about the effect of the explanatory variable
```

rel	Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
1	2.022222	1.29	0.2562	0.58471	6.99382
2	2.252252	7.69	0.0056	1.24857	4.06278
3	1.393519	0.79	0.3745	0.66775	2.90811
4	2.019724	2.15	0.1425	0.77414	5.26941
.	.	.	.	.	.

Mantel-Haenszel estimate controlling for rel

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
1.914248	10.89	0.0010	1.292931	2.834138

Test of homogeneity of ORs (approx): chi2(3) = 1.03
Pr>chi2 = 0.7931

The estimate of the odds ratio for **ed2** adjusted for religion is 1.91 (indicating substantial confounding by religion).

12.

```
xi:logistic case i.ed2
```

case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ied2_2	2.416169	.4492301	4.74	0.000	1.678287	3.478471

Logistic regression produces exactly the same estimate of the odds ratio as one obtains using simple techniques (cross product ratio).

```
xi:logistic case i.ed2 i.rel
```

case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ied2_2	1.922676	.3827857	3.28	0.001	1.301489	2.84035
_Irel_2	.5327553	.1716189	-1.95	0.051	.283352	1.001681
_Irel_3	.4660027	.1573667	-2.26	0.024	.2404043	.9033058
_Irel_4	.2168593	.082612	-4.01	0.000	.1027807	.4575564

Logistic regression produces an adjusted odds ratio (1.92) which is very close to that obtained with the Mantel-Haenszel approach (1.91)

Summary of main points

1. 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`, `area`, and `sex` as explanatory variables produces odds ratios for these three variables adjusted for the other two. We have assumed that there are no interactions in any of the models in this session. We shall cover models with interactions in the next session.

2. We can use the Wald test provided routinely with the logistic regression output to test the null hypothesis that a single crude or adjusted odds ratio is equal to 1. We can also use the LRT to test whether a variable is associated with the outcome after adjusting for the other variables in the model. This is particularly useful for categorical variables with more than two levels. **The Wald test and LRT are not tests of confounding – they are testing whether variables are associated with the outcome** (only part of the confounding story).

3. In order to determine whether a variable is a confounder, compare the crude (unadjusted) odds ratio with the adjusted odds ratio.

4. Logistic regression provides valid odds ratio estimates for unmatched or frequency matched case-control studies, but will not provide valid estimates of the baseline odds (or any other odds.

Summary of STATA commands

1. The LRT to estimate the effect of area adjusted for age is obtained by comparing the model with area (i.e. agegrp and area) with the model without area (i.e. agegrp):

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

2. The LRT must be performed on nested models, that is, where one model is a special case of the other – often this is when the variables in one model are a subset of the variables in the other model. For example, one can compare the model with agegrp, area, and sex (`xi: logistic mf i.agegrp i.area i.sex`) with the model with just sex (`xi: logistic mf i.sex`), because these models are nested. One cannot compare the model with agegrp and area (`xi: logistic mf i.agegrp i.area`) with the model with sex (`xi: logistic mf i.sex`) because these models are not nested.

3. In addition to being nested, 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. For example, in the Mwanza dataset, one observation had religion missing but no observations had missing values for age or education. Therefore, the log likelihood from a model of age and education is based on all 763 observations, whereas the log likelihood from a model of age, education, and religion is based on 762 observations. The following STATA commands will give a warning indicating that the LRT is not correct:

```
xi: logistic mf i.age2 i.ed2 i.rel
estimates store A
logistic mf i.age2 i.ed2
estimates store B
lrtest B A
```

The way to address this problem is to fit both models on the 762 observations which do not have any missing values:

```
xi: logistic mf i.age2 i.ed2 i.rel
estimates store A
```

```
logistic mf i.age2 i.ed2 if rel!=.
estimates store B
lrtest B A
```

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 11 SOLUTIONS

1. The model with **area** and **agebin** as in Section 2 of the lecture notes is:

```
. xi: logistic mf i.area*i.agebin
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]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.416252   .5418676     3.93   0.000     1.556869    3.750009
  _Iagebin_1 |    5.80094   1.236674     8.25   0.000     3.819749    8.809715
_IareXage_~1 |   1.592047   .4481527     1.65   0.099     .9169535    2.764169
------------------------------------------------------------------------------
```

i) the (stratum-specific) odds ratio for area in the baseline age group (i.e. in the 0-19 age group) is 2.41.

ii) the (stratum-specific) odds ratio for age group 1 vs 0 in the baseline group of area (ie. in the savannah) is 5.80.

iii) the interaction term (odds ratio scale) is 1.59

To obtain an estimate of the odds ratio for age in the forest area we need to combine the odds ratio for age in the baseline (savannah) area with the interaction parameter;

i.e. OR = 5.80 × 1.59 = 9.24.

You will notice this is the same as the ratio of odds in these groups from the tables at the beginning of Section 2 of the notes i.e. 6.677/0.723, because the interaction term has allowed us to model independently all 4 cells of this table.

2. The output from the model with an **area** and **agegrp** as in section 3 of the lecture notes is as follows:

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

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   1.827532   .6369638     1.73   0.084     .9229731    3.618602
  _Iagegrp_1 |     2.1175   .7949557     2.00   0.046     1.014527    4.419601
  _Iagegrp_2 |   6.963971   2.150749     6.28   0.000     3.801652    12.75679
  _Iagegrp_3 |       10.5   3.353467     7.36   0.000     5.614802     19.6356
_IareXage_~1 |   1.387791    .642613     0.71   0.479     .5599863    3.439305
_IareXage_~2 |   1.663802   .6901372     1.23   0.220     .7379504     3.75125
_IareXage_~3 |    2.58819   1.133702     2.17   0.030     1.096845     6.10727
------------------------------------------------------------------------------
```

i) The odds ratio for the effect of area in age group 0 is 1.83.

ii) The odds ratios for effects of age in the savannah are, respectively, 2.11, 6.96 and 10.50 for age groups 1, 2 and 3 versus age group 0.

iii) The parameters in bold are the interaction terms which show how the effect of area (the odds ratio) differs in age groups 1, 2 & 3 from that in age group 0 (1.39, 1.66, 2.59).

Age specific ORs for the effect of area are derived by combining the odds ratio for area at the baseline of **agegrp** with the appropriate interaction parameter for each level of **agegrp.**

Age category	OR (forest vs savannah)
Agegroup 0	1.83 from **_Iarea_1** parameter
Agegroup 1	1.83 × 1.39=2.54 from **_Iarea_1** and **_IareXage_1_1** parameters
Agegroup 2	1.83 × 1.66=3.04 from **_Iarea_1** and **_IareXage_1_2** parameters
Agegroup 3	1.83 × 2.59=4.73 from **_Iarea_1** and **_IareXage_1_3** parameters

vi)

```
. lincom _Iarea_1 + _IareXage_1_1,or
 ( 1)  _Iarea_1 + _IareXage_1_1 = 0.0
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.536232   .7731736     3.05   0.002     1.395401    4.609766
------------------------------------------------------------------------------

. lincom _Iarea_1 + _IareXage_1_2,or
 ( 1)  _Iarea_1 + _IareXage_1_2 = 0.0
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    3.04065   .6838193     4.94   0.000     1.956761    4.724929
------------------------------------------------------------------------------

. lincom _Iarea_1 + _IareXage_1_3,or
 ( 1)  _Iarea_1 + _IareXage_1_3 = 0.0
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |        4.73   1.254928     5.86   0.000     2.812074    7.956015
------------------------------------------------------------------------------
```

3.

	LR test for interaction: $\chi 2$; df	p-value
agebin	2.70	0.10
agegrp	5.27	0.153

The p-value for the interaction term with the binary age variable is smaller. This may reflect increased power to detect interaction when fewer age groups and hence fewer parameters are added to the model. (see section 7 of the lecture notes).

In the models with interaction terms the parameter **_Iarea_1** represents a stratum-specific odds ratio - the odds ratio for area in the youngest (baseline) age group. In the models without interaction terms it represents a summary adjusted odds ratio – the odds ratio for area adjusted for age (like a Mantel-Haenszel summary odds ratio).

4.

```
. xi:logistic mf i.agegrp i.area*i.agebin
Logistic regression                               Number of obs   =       1302
                                                  LR chi2(5)      =     332.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -690.83482                       Pseudo R2       =     0.1939

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagegrp_1 |   2.630939   .5771984     4.41   0.000      1.71146    4.044407
  _Iagegrp_2 |   7.295581   1.893735     7.66   0.000     4.386429    12.13413
  _Iagegrp_3 |   13.07112   3.511137     9.57   0.000     7.720847    22.12895
    _Iarea_1 |   2.206122   .5067814     3.44   0.001     1.406356    3.460701
_IareXage_~1 |   1.667168   .4778762     1.78   0.075     .9505861    2.923933
------------------------------------------------------------------------------
```

The odds ratio for the effect of area in those aged 0-19 (baseline of **agebin**) is 2.206. The odds ratio for the effect of area in those aged 20+ is $2.206 \times 1.667 = 3.68$.

```
. estimates store A
. xi:logistic mf i.agegrp i.area
Logit estimates                                   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]
-------------+----------------------------------------------------------------
  _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
    _Iarea_1 |   3.083224    .424372     8.18   0.000     2.354217    4.037975
------------------------------------------------------------------------------

. estimates store B
. lrtest B A
likelihood-ratio test                             LR chi2(1)  =       3.15
(Assumption: B nested in A)                       Prob > chi2 =     0.0762
```

The evidence for an interaction between age and area is still not strong. The data are not incompatible with the hypothesis of no interaction.

5.

```
. xi: logistic mf i.agegrp i.sex i.area
Logit estimates                                   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]
-------------+----------------------------------------------------------------
  _Iagegrp_1 |   2.567357   .5748449     4.21   0.000      1.65538    3.981758
  _Iagegrp_2 |   10.46237   2.205354    11.14   0.000     6.921597    15.81446
  _Iagegrp_3 |   17.65935   3.834073    13.22   0.000     11.53899    27.02598
     _Isex_1 |   .5591696   .0758462    -4.29   0.000     .4286332    .7294595
    _Iarea_1 |   3.073138    .425981     8.10   0.000     2.342034    4.032467
------------------------------------------------------------------------------
```

The estimated odds ratio for the effect of area, controlling for age and sex (and assuming no interaction between age and sex) is 3.07 with a 95% c.i. (2.34 to 4.03). There is strong evidence for an association between area and odds of infection after controlling age and sex (Wald test p<0.001)

```
. xi: logistic mf i.agegrp*i.sex i.area
Logistic regression                               Number of obs   =       1302
                                                  LR chi2(8)      =     352.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -680.61501                       Pseudo R2       =     0.2058

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagegrp_1 |   3.070077   .9595522     3.59   0.000      1.66382      5.6649
  _Iagegrp_2 |   12.45813   3.892053     8.07   0.000     6.753489     22.98144
  _Iagegrp_3 |    14.5496   4.357702     8.94   0.000     8.089296     26.16927
     _Isex_1 |   .6020671   .2080773    -1.47   0.142     .3058198     1.185289
_IageXse~1_1 |   .6897617   .3105214    -0.83   0.409       .28543     1.666858
_IageXse~2_1 |   .7525774   .3159721    -0.68   0.498        .3305     1.713685
_IageXse~3_1 |   1.434654   .6136086     0.84   0.399     .6204176     3.317495
    _Iarea_1 |   3.059568   .4254645     8.04   0.000     2.329656     4.018171
------------------------------------------------------------------------------
```

Allowing for interaction between age and sex has not changed to any important extent our estimate of the effect of area (now 3.06) or our assessment of its statistical significance. Therefore, we might as well stick with the simpler model which does not include an interaction between age and sex.

Key points

- Unless you explicitly include interaction terms in your logistic regression model, Stata (and other packages) assume no interactions between any of the variables. I.e. the default is to assume no interaction.
- You should check that this assumption is reasonable.
- Interaction terms (and stratum-specific odds ratios) should be examined.
- The assumption/hypothesis of no interaction can be tested formally using the likelihood ratio test.
- We may be able to increase the power of this test by combining groups.
- The measures of interest, stratum-specific effects, can be obtained by multiplying together the relevant parameter estimates (baseline estimates and interaction estimates). Confidence intervals for the stratum-specific estimates (and the estimates themselves) can be obtained using the `lincom` command.
- complex interaction terms between confounders often have little effect on the estimate of the effect of the main exposure.
- The interpretation of parameters in models with and without interactions is different. When interactions are involved we obtain stratum-specific odds ratios. When interactions are not involved we obtain summary, adjusted odds ratios.

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 12 SOLUTIONS

1 a)

```
use onch667b
tab lesions agegrp, col
```

Lesions present	Age group 0	1	2	3	Total
0	185	142	92	59	478
	88.52	72.08	60.93	53.64	71.66
1	24	55	59	51	189
	11.48	27.92	39.07	46.36	28.34
Total	209	197	151	110	667
	100.00	100.00	100.00	100.00	100.00

There appears to be a strong trend in the percentage of people with eye lesions by age, from 11% in the youngest to 46% in the oldest.

b) To calculate the log odds of eye lesions by age group:

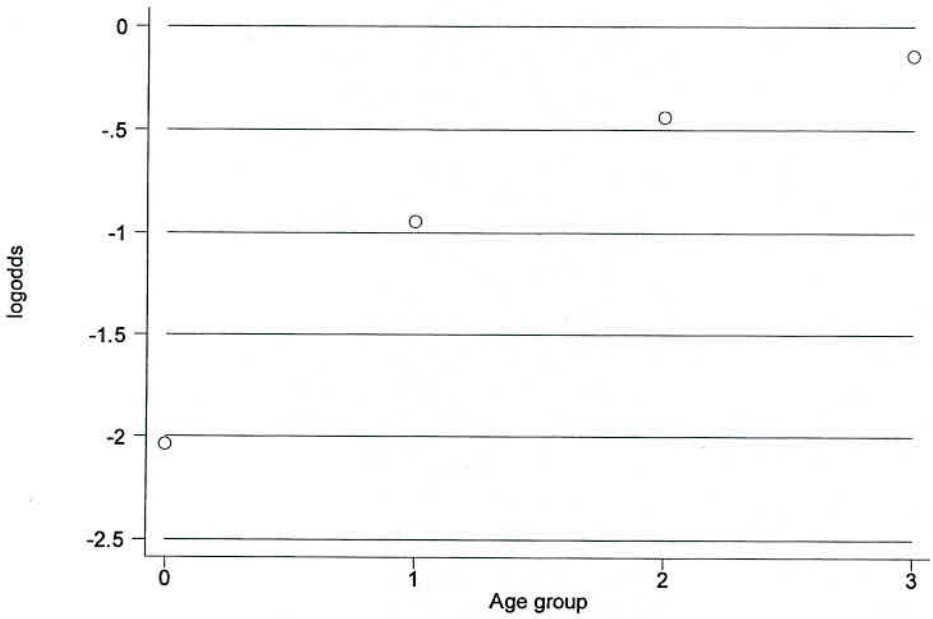
```
tabodds lesions agegrp
```

agegrp	cases	controls	odds	[95% Conf.	Interval]
0	24	185	0.12973	0.08479	0.19848
1	55	142	0.38732	0.28372	0.52877
2	59	92	0.64130	0.46248	0.88927
3	51	59	0.86441	0.59425	1.25738

```
Test of homogeneity (equal odds): chi2(3)  =    55.34
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    53.64
                                  Pr>chi2  =   0.0000
```

$\log_e(24/185) = -2.042$
$\log_e(55/142) = -0.948$
$\log_e(59/92) = -0.444$
$\log_e(51/59) = -0.146$

c)



d) An increase of 20 years corresponds to an increase of 2 in the value of **agegrp**. The log(odds) increase by 1.1, 0.5 and 0.3 at each step. You might have estimated a change of around 1.2.

e) Logistic regression of the log(odds) of eye lesions by age group:

```
. logistic lesions agegrp

Logistic regression                               Number of obs   =        667
                                                  LR chi2(1)      =      53.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -370.63346                       Pseudo R2       =     0.0678

------------------------------------------------------------------------------
     lesions | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   1.827737    .155406     7.09   0.000     1.547174    2.159178
------------------------------------------------------------------------------
```

```
. logit

------------------------------------------------------------------------------
     lesions |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   .6030787   .0850265     7.09   0.000     .4364299    .7697275
       _cons |  -1.762182   .1567976   -11.24   0.000      -2.0695   -1.454865
------------------------------------------------------------------------------
```

f) The slope of the regression line is 0.603. This means that, on average, the log(odds) of eye lesions increase by 0.603 for a unit increase in age group. So for an increase of 20 years the log odds change by an estimated 1.206

g) Likelihood ratio test comparing age group as a categorical variable with age group as a quantitative variable:

```
. xi:logistic lesions i.agegrp
Logit estimates                                   Number of obs   =        667
                                                  LR chi2(3)      =      58.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -368.15413                       Pseudo R2       =     0.0741

------------------------------------------------------------------------------
     lesions | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagegrp_1 |   2.985622   .8027723     4.07   0.000     1.762643    5.057145
  _Iagegrp_2 |   4.943388   1.352816     5.84   0.000     2.891234    8.452127
  _Iagegrp_3 |   6.663136   1.926892     6.56   0.000     3.780266    11.74451
------------------------------------------------------------------------------

. estimates store A

. logistic lesions agegrp

Logit estimates                                   Number of obs   =        667
                                                  LR chi2(1)      =      53.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -370.63346                       Pseudo R2       =     0.0678

------------------------------------------------------------------------------
     lesions | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   1.827737    .155406     7.09   0.000     1.547174    2.159178
------------------------------------------------------------------------------

. estimates store B
. lrtest B A
likelihood-ratio test                             LR chi2(2)  =      4.96
(Assumption: B nested in A)                       Prob > chi2 =    0.0838
```

There is weak evidence that the model with age group as a categorical variable fits better than the model with age group as a linear effect (i.e. weak evidence against the null hypothesis that the association between age group and the log odds of disease is linear).

h) (optional question)

Test of the null hypothesis that the effect of age group is linear using a quadratic effect:

```
generate agegrp2=agegrp*agegrp

logistic lesions agegrp agegrp2

Logistic regression                               Number of obs   =        667
                                                  LR chi2(2)      =      58.62
                                                  Prob > chi2     =     0.0000
Log likelihood = -368.28536                       Pseudo R2       =     0.0737

------------------------------------------------------------------------------
     lesions | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   3.346516   .9960464     4.06   0.000     1.867429    5.997108
     agegrp2 |   .8201915    .075579    -2.15   0.031      .684666    .9825434
------------------------------------------------------------------------------
```

```
logit
```

```
------------------------------------------------------------------------------
     lesions |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |     1.20792   .2976368     4.06   0.000      .6245625    1.791277
     agegrp2 |    -.1982175    .092148    -2.15   0.031     -.3788242   -.0176108
       _cons |    -2.010522    .205394    -9.79   0.000     -2.413087   -1.607958
------------------------------------------------------------------------------
```

```
estimates store A
```

```
logistic lesions agegrp
```

```
Logit estimates                                   Number of obs   =        667
                                                  LR chi2(1)      =      53.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -370.63346                       Pseudo R2       =     0.0678

------------------------------------------------------------------------------
     lesions | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |   1.827737    .155406     7.09   0.000     1.547174    2.159178
------------------------------------------------------------------------------
```

```
estimates store B
lrtest B A
Logistic:  likelihood-ratio test                  chi2(1)     =        4.70
                                                  Prob > chi2 =      0.0302
```

The evidence against the null hypothesis (linear trend in log(odds)) is stronger than previously. This is likely to be because we have used fewer degrees of freedom in the test (1 this time compared with 2 previously), and the alternative hypothesis we have specified (quadratic relationship) looks reasonable.

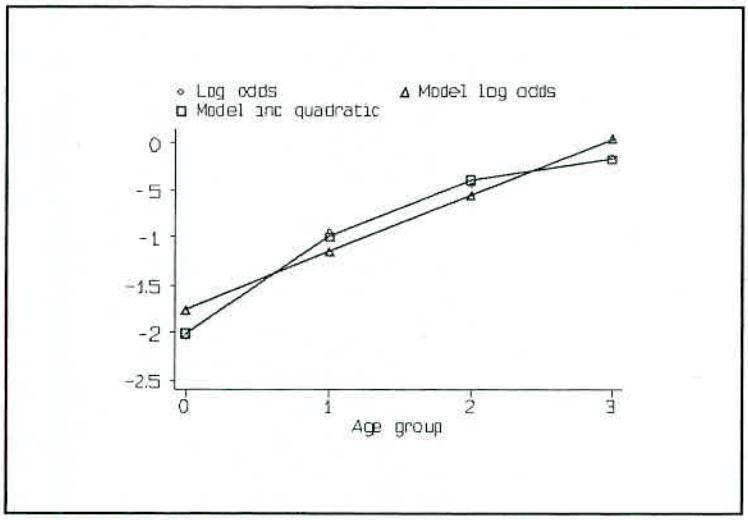
The coefficients can be used to calculate the fitted log(odds) from the quadratic model, as follows:

Age group			**Fitted log(odds)**
0			-2.011
1	-2.011 + 1.208 - .1982	=	-1.001
2	-2.011 + 2×1.208 - 4×.1982	=	-0.388
3	-2.011 + 3×1.208 - 9×.1982	=	-0.171

These values can also be derived by using the `predict` command after fitting a logistic regression model. This gives the predicted probability of disease for each observation, which can be converted into the odds or log odds.

```
predict lp
generate logo=log(lp/(1-lp))
tab agegrp logo
```

```
           | logo
 Age group | -2.010522   -1.00082  -.3875526  -.1707201 |     Total
-----------+--------------------------------------------+----------
         0 |       209          0          0          0 |       209
         1 |         0        197          0          0 |       197
         2 |         0          0        151          0 |       151
         3 |         0          0          0        110 |       110
-----------+--------------------------------------------+----------
     Total |       209        197        151        110 |       667
```



2a)

```
use onchall

xi:logistic mf i.area agegrp

Logistic regression                                     Number of obs =   1302
                                                        chi2(2)       = 322.47
Log Likelihood =  -695.7927                             Pseudo R2     = 0.1881

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.930859    .394344     7.99   0.000     2.251473    3.815252
      agegrp |   2.681743   .1808225    14.63   0.000     2.349757    3.060633
------------------------------------------------------------------------------
```

The odds ratio for area is 2.93. When **agegrp** was specified as a categorical variable, the odds ratio was 3.08. The crude odds ratio was 2.41. It appears that using the categorical variable in this case slightly improves the control of confounding by age compared to using a continuous variable.

b) *(optional question)*
Deriving the linear interaction between agegrp and area:

```
xi:logistic mf i.area*agegrp
i.area            _Iarea_0-1          (naturally coded; _Iarea_0 omitted)
i.area*agegrp     _IareXagegr_#       (coded as above)

Logistic regression                               Number of obs   =       1302
                                                  LR chi2(3)      =     328.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -692.56131                       Pseudo R2       =     0.1919
```

```
------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   1.616517    .429584     1.81   0.071     .9602327    2.721349
      agegrp |   2.223374   .2138813     8.31   0.000     1.841322    2.684697
_IareXageg~1 |   1.409155   .1895728     2.55   0.011     1.082548      1.8343
------------------------------------------------------------------------------
```

```
. estimates store A

. xi:logistic mf i.area agegrp
i.area            _Iarea_0-1          (naturally coded; _Iarea_0 omitted)

Logit estimates                           Number of obs   =       1302
                                          LR chi2(2)      =     322.47
                                          Prob > chi2     =     0.0000
Log likelihood =  -695.7927               Pseudo R2       =     0.1881

------------------------------------------------------------------------------
          mf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iarea_1 |   2.930859    .394344     7.99   0.000     2.251473    3.815252
      agegrp |   2.681743   .1808225    14.63   0.000     2.349757    3.060633
------------------------------------------------------------------------------

. estimates store B
. lrtest B A
likelihood-ratio test                     LR chi2(1)  =      6.46
(Assumption: B nested in A)               Prob > chi2 =    0.0110
```

There is reasonably strong evidence that the odds of infection increase more steeply with age in the forest than in the savanna

3. Association between HIV infection and the number of infections in the past year (MWANZA dataset):

```
. use mwanza
. tab case inj
```

Case/contro| l	Injections in past 1 year 1	2	3	4	5	9	Total
0	278	53	124	83	35	1	574
1	68	15	41	41	23	1	189
Total	346	68	165	124	58	2	763

The two missing values for `inj` should be recoded to missing:

```
. recode inj 9=.
(2 changes made)

. tab case inj, col
```

Case/contro| l	Injections in past 1 year 1	2	3	4	5	Total
0	278	53	124	83	35	573
	80.35	77.94	75.15	66.94	60.34	75.30
1	68	15	41	41	23	188
	19.65	22.06	24.85	33.06	39.66	24.70
Total	346	68	165	124	58	761
	100.00	100.00	100.00	100.00	100.00	100.00

The column percentages shown in the table must not be interpreted as risks as we are now dealing with a case control study. However, the pattern, with increased percentages infected with increasing numbers of injections in the past year, is informative.

tabodds case inj

inj	Cases	Controls	_odds	95%	Conf. Interval
1	68	278	0.245	0.188	0.319
2	15	53	0.283	0.160	0.502
3	41	124	0.331	0.232	0.471
4	41	83	0.494	0.340	0.718
5	23	35	0.657	0.388	1.112

```
Test of homogeneity (equal odds) :  chi2(4) = 16.61
                                    pr>chi2 = 0.0023
Score test for trend of odds        chi2(1) = 15.36
                                    pr>chi2 = 0.0001
```

Again, the odds shown should not be interpreted literally as this is a case control study, but the hypothesis test results are valid. There is strong evidence that the odds of HIV infection increase with increasing numbers of injections.

logistic case inj

```
Logit estimates                                   Number of obs   =        761
                                                  LR 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     3.89   0.000     1.122694    1.421009
------------------------------------------------------------------------------
```

This analysis included the zero group. Now repeat the analysis excluding individuals who had no injections in the past year:

logistic case inj if inj!=1

```
Logit estimates                                   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
------------------------------------------------------------------------------
```

logit

```
------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inj |   .3084991   .1188025     2.597   0.009      .0756505    .5413477
       _cons |   -1.96961   .4324073    -4.555   0.000     -2.817112   -1.122107
------------------------------------------------------------------------------
```

Although the z statistic (Wald test) is somewhat reduced after excluding the zero group, there is still strong evidence of a trend for increasing odds of HIV infection with increasing number of injections in the past year.

Note: This result is very difficult to interpret. One reason for injections may be because the subject had an HIV-related illness. It does **not** prove that injection is a major mode of transmission of HIV in Mwanza!

Key points:

- fitting a quantitative exposure as a linear trend may be more appropriate than fitting separate categories if this does not significantly reduce the fit of the model (i.e. the likelihood).

- the estimate for a continuous variable is interpreted as the increase in log odds for a unit increase in the variable. Remember this may be an increase of one category, for an ordered categorical variable, or an increase of one unit if the variable is on its original scale (such as age in years).

STATISTICAL METHODS IN EPIDEMIOLOGY

SESSION 14 PRACTICAL SOLUTIONS

Analysis of Matched Case-Control Studies

Individually matched studies

1.

```
. use diabraz
. match case bf pair
```

1 bf	0 bf 1	2	Total
1	24	6	30
2	29	27	56
Total	53	33	86

```
. mhodds case bf pair, c(1,2)

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

Note: only 35 of the 86 strata formed in this analysis
contribute information about the effect of the explanatory variable
```

Odds ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
0.206897	15.11	0.0001	0.085900	0.498327

These answers are the same as those obtained in the lecture. Note STATA's reminder that only the discordant pairs contribute to the analysis.

2. `. match case wat2 pair`

1 wat2	0 wat2 1	2	Total
1	56	2	58
2	11	17	28
Total	67	19	86

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

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
```

Odds ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
5.500000	6.23	0.0126	1.219099	24.813414

Those with no water supply in the house/plot (wat2=2) have a higher risk of diarrhoea than those with water supply(wat2=1): an OR of more than 5.

`. match case bwtgp pair`

1 bwtgp	0 bwtgp 1	2	Total
1	31	18	49
2	25	12	37
Total	56	30	86

```
. mhodds case bwtgp pair, c(2,1)

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

Note: only 43 of the 86 strata formed in this analysis
contribute information about the effect of the explanatory variable
```

Odds ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
1.388889	1.14	0.2858	0.757781	2.545608

Those with lower birth weight (bwtgp=2)have a higher risk than those with higher birth weight.

Unmatched analyses of **wat2** and **bwtgrp** produce the following estimates:

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

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

------------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   1.702359       2.36        0.1247         0.856308    3.384327
------------------------------------------------------------------------

. mhodds case bwtgp, c(2,1)

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

------------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------
   1.409524       1.19        0.2751         0.758640    2.618840
------------------------------------------------------------------------
```

Ignoring the matching in the analysis results in a substantial *underestimate* of the odds ratio for `wat2` (1.70 instead of 5.50). This is because water supply is strongly associated with the matching factor (neighbourhood). So matching on neighbourhood makes the cases and controls more similar with respect to water supply than they would have been in a non-matched design. Matching makes the controls less likely to have a water supply. If we then ignore the matching we underestimate the effect of water supply.

For `bwtgrp`, ignoring the matching makes little difference (1.41 vs 1.39). This suggests that birthweight is not strongly associated with either of the matching factors (neighbourhood and age).

3. In the younger age group, the odds ratio is $2/9 \approx 0.22$. In the older age group it is 4/20=0.2.

Optional The table in the notes can be obtained as follows (NB the lines starting with '*' are comments and don't have to be typed)

```
gen agegp4=age
recode agegp4 0/2=1 3/max=2
* check the new variable is as we want it:
table age agegp4, missing

* do the subtables one at a time (by: doesn't work with match)
match case bf pair if agegp4==1
match case bf pair if agegp4==2
```

The 2x2 table for the interaction can be obtained as follows:

	Age group	
	0-2	3+
Case +, control -	2	4
Case -, control +	9	20

We can put these into a chi-squared test using the `cci` command:

```
cci 2 4 9 20
```

Which yields chi-squared = 0.01, p=0.91. So there is no evidence that the effect of breast feeding varies with age.

Note that we are using this command simply as convenient way to get a chi-squared test from a set of cell frequencies (as opposed to long variables). We are only interested in the chi-squared statistic & p value, not the odds ratios etc, which do not apply to the way we are using the command here.

4.

```
. clear
. use diabraz2
. mhodds case bf set, c(1,2)

Mantel-Haenszel estimate of the odds ratio
Comparing bf==1 vs bf==2, controlling for set

Note: only 131 of the 170 strata formed in this analysis
contribute information about the effect of the explanatory variable

------------------------------------------------------------------
 Odds ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------
   0.246377      41.28        0.0000         0.155045   0.391510
------------------------------------------------------------------
```

Applying the formula in the notes gives MHOR=$(n_{11}+2n_{10})/(2n_{02}+n_{01})$= (24+10)/(72+66)=0.246, the same as from `mhodds`. Only those triplets in which the case and both controls were unexposed (n_{00}=18), or in which all three were exposed (n_{12}=21) do not contribute. Hence the message from `mhodds` that 131 (=170-(18+21)) sets contribute.

Optional for those curious about how to get the table in the practical notes from STATA. The following commands will do it, although there may be an easier way! NB the lines starting with * are just comments and don't have to be typed.

```
use diabraz2, clear

preserve
* recode the bf field to a new one called newbf,
* in which 0 means not breast fed
gen newbf=bf
recode newbf 2=0
* count up the number breast fed per set
* (whether cases or controls)
egen conbf=sum(newbf),by(set)
* drop the controls
keep if case==1
* count the number of controls who were breast fed:
* this is the total number (in newbf),
* minus the number of cases breast fed
* (which is 0 or 1, held in newbf)
replace conbf=conbf-newbf
rename newbf casebf
tab casebf conbf
restore
```

5. To repeat Q1:

```
. clear
. use diabraz
. xi: clogit case i.bf, str(pair)
```

Note that the results are on the log scale, and are the reciprocals of those in Q1.

Q2 can be done similarly:

```
. xi: clogit case i.wat2, str(pair)
. xi: clogit case i.bwtgp, str(pair)
```

For Q4:

```
. clear
. use diabraz2
. xi: clogit case i.bf, str(set)
```

NB in this example, `clogit` does not give exactly the same odds ratio (3.8) as `mhodds` in Q4 (4.1). This is because, in general, the maximum likelihood and Mantel-Haenszel estimates are not equal, although they will be similar. For one-to-one matching they *are* equal, but not for one-to-two. (Similarly, in stratified analysis of unmatched case-control studies, they will, in general, not be exactly equal.)

Frequency matched studies

6. (a) Unmatched study

Males	Exposed	Unexposed	Total
Cases	450	10	460
Controls	270	30	300

Females	Exposed	Unexposed	Total
Cases	50	90	140
Controls	30	270	300

Mantel-Haenszel OR = 5.00, with a 95% CI of 3.22-7.78

(b) Frequency Matched study

Males	Exposed	Unexposed	Total
Cases	450	10	460
Controls	414	46	460

Females	Exposed	Unexposed	Total
Cases	50	90	140
Controls	14	126	140

Mantel-Haenszel OR = 5.00, with a 95% CI of 3.04-8.29

These are not the only correct tables. In this example, all cases were included, and an equal number of controls. Other sampling fractions are possible, e.g. twice as many controls as cases. However, to help comparability of the two designs, the matched study should have the same number of cases as the unmatched one, and similarly for controls.

The matched study has a *wider* confidence interval than the unmatched study. This is because the matching variable (sex) is associated with exposure but is not a risk factor for disease.

(Optional, more advanced, point: So we would not expect to gain precision by matching on sex, but why do we *lose* precision? We can see this algebraically by looking back to Session 4 for the formula for the variance of log(OR): $1/D_1 + 1/D_0 + 1/H_1 + 1/H_0$, i.e. the sum of the reciprocal cell frequencies. This is dominated by the smallest of the four cells. For males, the matched and unmatched tables both have the same smallest cell – 10 unexposed cases. However, for females, the matched table has a much smaller smallest cell: 14 as opposed to 30.)

For the unmatched design, an unstratified analysis uses the following table:

	Exposed	Unexposed	Total
Cases	500	100	600
Controls	300	300	600

The OR for this is again 5.0, with a 95% CI of 3.79-6.60.

If we (wrongly) perform an unmatched (unstratified) analysis of the matched study we obtain the following table:

	Exposed	Unexposed	Total
Cases	500	100	600
Controls	428	172	600

The estimate of the odds ratio is now 2.01. So for the matched study a (sex) stratified analysis produces an OR of 5.0 while a crude (unstratified) analysis produces an OR of 2.01. There is a big difference between the crude and adjusted estimates of the OR indicating that sex is a confounder in the matched study, even though it wasn't in the population as a whole. In this situation, matching has *introduced* confounding where none existed. An unmatched analysis of the unmatched study produces an unbiased estimate of the OR (=5.0) because sex is not a confounder in this population.

Optional – individually matched studies

7. This is a somewhat open-ended question (each person should do as much as they want) so there are not specific answers for all the possible analyses. The main point of this session is the classical Mantel-Haenszel techniques. The conditional logistic regression is here really to show you where to start if you have to analyse such data in the future.

   For confounding, e.g. with social class on the effect of breast feeding, do

```
. xi: clogit case i.bf i.social, str(set)
```

   and compare with results from the model with breast feeding alone. For this example, the unadjusted OR for breastfeeding is 3.79 and the adjusted one is 3.70, so not much confounding by social class.

Key points

Matched design implies matched analysis: summarize the data using tables in which pairs are the units, not individuals.

Only the discordant pairs contribute to the analysis. This number can be much fewer than the total number of pairs. This should be borne in mind when designing a case-control study.

epilepsy. This is the first matched case-control study to be carried out on this theme. Indeed, apart from only one previous case-control study (9), all of the previous studies have been cross-sectional (5-8). This matched case-control study in the Central African Republic did not show any relation between *O. volvulus* infestation and epilepsy. These results are in concordance with those of Kaboré et al. (7), who had examined 1,046 subjects in Burkina Faso who were above age 15. The prevalence of onchocerciasis in this region was 12.9 percent, and that of epilepsy was 1.5 percent; only two epileptics had onchocerciasis. This study, performed in an onchocerciasis hypoendemic area, where aerial fumigation and ivermectin mass treatment have been ·ıg on for several years now, had, like the present ıy, shown absolutely no relation between onchocerciasis and epilepsy.

However, several past studies have favored this relation. One study was done in Kyarusozi (Uganda) (5) in an onchocerciasis hyperendemic zone, where 231 persons were examined. Among the subjects who had onchocerciasis, 61 percent had epilepsy and 70 percent had growth retardation. The level of prevalence of epilepsy was 2.0 percent, but 91 percent of the population was below age 19 years. Another study was carried out in Uganda (6) in which authors compared the prevalence of epilepsy between two villages, one situated in an onchocerciasis hyperendemic zone and the other in an hypoendemic zone. In the first village, 8 percent of the subjects were classified as epileptics, and only 0.2 percent in the second village were. The relative risk for developing epilepsy, adjusted for age, sex, and ethnic background, was 6.5 times higher for persons with onchocerciasis (95 percent confi- ·ıce interval 3-15). These results were, however, disputed (28), since there seemed to be a high relation between the village and epilepsy, rather than between onchocerciasis and epilepsy. Other unexplained cofactors, such as isolated familial epilepsy, obstetric problems, nutritional, cerebral infections, or other parasitic infections, could probably be responsible for the epilepsy. Still, in the same country (Uganda), Kaiser et al. (8) carried out a study on 4,743 subjects in the parish of Kabende, district of Kabarole. The prevalence of onchocerciasis was determined in each of the 13 villages concerned (sampling about 30 subjects per village). The prevalence of onchocerciasis varied between 15 and 85 percent, depending upon the village. Sixty-one subjects (1.3 percent) were confirmed epileptics. There was a positive correlation between the prevalence of epilepsy and endemicity of onchocerciasis. Newell et al. (9) had, in a case-control study, examined 110 epileptics patients and 82 controls in two administrative zones with different endemicity for onchocerciasis (mesoendemic and hyperendemic). Epilepsy was more prevalent in the onchocerciasis hyperendemic zone. In the mesoendemic area, the difference in prevalence of onchocerciasis between epileptics and controls was not significant. In this study, there were fewer controls than epileptic cases. Moreover, the controls had received ivermectin more often than had the epileptics. Therefore, some controls could have been classified as negative for onchocerciasis, which could explain the difference between the prevalences of onchocerciasis among patients and controls.

It is difficult to establish a link between epilepsy and onchocerciasis. The methodologies and conclusions of the epidemiologic studies quoted in the literature differ. The etiologies of epilepsies are many and are frequently related to the sequelae of head injuries, neonatal head traumas, and infections. Epilepsy is described as a complication in other lymphatic filariasis, such as *Wuchereria bancrofti* or *Mansonella perstans*. The embryos of these filariae live in their adult stage within the vessels and may migrate anywhere in the body through the blood and thus penetrate the central nervous system (29). The neurologic complications may occur due to wandering of microfilariae in the brain tissue or meningeal spaces or to migration of adult worms in the neuromeningial spaces (30). The discovery of *O. volvulus* microfilariae in the cerebrospinal fluid by Mazzotti (31) was certainly accidental and possibly related to the cutaneous abrasion by the needle during the lumbar puncture, taking with it the microfilariae. In the onchocerciasis, there exists no microfilaraemia. The filariae live in the adult state buried in the dermis, preferably on the bony prominences and the microfilariae migrate subdermally, and thus, only adult filariae may end up in the central nervous system. Indeed, if any cause-effect relation exists between onchocerciasis and epilepsy, it is most certainly not a direct mechanical effect, and other pathophysiologic mechanisms must be explored.

ACKNOWLEDGMENTS

This study was partially financed by the "Ministère délégué à la Coopération et à la Francophonie; Ministère des Affaires étrangère" Contract CAMPUS reference 96.413.101, Committee of the 17/9/96 decisions FAC/IG/94.0164.00 and 91.0198.00 and "Conseil Régional du Limousin."

The authors thank the Ministry of Public Health and Population of Central African Republic; the Faculty of Health Sciences, University of Bangui (Central African Republic); the National Programme for the Fight Against Onchocerciasis and Blindness (Bossangoa, Central African

Republic); Dr. Ione Bertocci, head of Ngaoundaye hospital (Central African Republic); the Deutsche Gessellschaft für Technische Zusammenarbeit of Bossangoa (Central African Republic); the Pasteur Institute of Bangui (Central African Republic); Professor François Denis, head of the Department of Virology of Limoges; Professor Marie-Laure Darde, head of the Department of Parasitology of Limoges; Dr. Philippe Gaxotte, Merck Sharp & Dohme Pharmaceutical Firm (Paris, France); the BioMérieux Laboratory (Craponne, France); and the Catholic missions of Bossangoa and Paoua (Central African Republic).

REFERENCES

1. Roblès R. Onchocercose humaine au Guatemala produisant la cécité et "l'érysipèle du littoral" (Erysipela de la costa). (In French). Bull Soc Path Exot 1919;2:442-60.
2. Puyelo R, Holstein M. L'onchocercose humaine en Afrique Noire Française. Maladie sociale. (In French). Med Trop (Mars) 1950;10:397-501.
3. Rwiza HT. Assessment of the size of the problem, organization of an epilepsy care system and research on the risk factors. Trop Geogr Med 1994;46:22-4.
4. Jilek-Aall L. Neurofilariasis: can onchocerciasis cause epilepsy? In: Rose C, ed. Recent advances in tropical neurology. Amsterdam, the Netherlands: Elsevier Science 1995:283-8.
5. Ovuga E, Kipp W, Mungherera M, et al. Epilepsy and retarded growth in a hyperendemic focus of onchocerciasis in rural western Uganda. East Afr Med J 1992;69:554-6.
6. Kipp W, Burnham G, Burnham J. Onchocerciasis and epilepsy in Uganda. Lancet 1994;343:183-4.
7. Kaboré JK, Cabore JW, Melaku Z, et al. Epilepsy in a focus of onchocerciasis in Burkina Faso. Lancet 1996;347:836.
8. Kaiser C, Kipp W, Asaba G, et al. Prevalence of epilepsy follows the distribution of onchocerciasis in a West Uganda focus. Bull World Health Organ 1996;74:361-7.
9. Newell ED, Vyungimana F, Bradley JE. Epilepsy, retarded growth and onchocerciasis, in two areas of different endemicity of onchocerciasis in Burundi. Trans R Soc Trop Med Hyg 1997;91:525-7.
10. Commission on Classification and Terminology of the International League Against Epilepsy. Proposal for revised clinical and electroencephalographic classification of epilepsies and epileptic syndromes. Epilepsia 1989;30:389-99.
11. Preux PM, Druet-Cabanac M, Zenebe M, et al. Limoges epilepsy protocol: research tool in tropical countries. Presented at the 12th Congress of the Pan African Association of Neurological Sciences, Durban, South Africa, May 19-22, 1996.
12. Schoënberg BS. Clinical neuroepidemiology in developing countries: neurology with few neurologists. Neuroepidemiology 1987;3:143-53.
13. Guerra G, Flisser A, Canedo L, et al. Biochemical immunological characterization of antigen B purified from cysticercosis of *Taenia solium*. In : Flisser A, Willms K, LaClette JP, et al., eds. Cysticercosis present state of knowledge and perspectives. New York, NY: Academic Press, Inc., 1982:437-51.
14. Chamouillet H, Bouteille B, Isautier H, et al. Séroprévalence de la cysticercose, teniasis et ladrerie porcine, à la réunion en 1992. (In French). Med Trop (Mars) 1997;57:41-6.
15. Cavallo AP. Enquête sur l'endémie onchocerquienne en RCA dans le foyer de l'Ouham-Pende. (In French). Bull OCEAC 1984;63:63-72.
16. Testa J, Feindirongai G, Auzemary A, et al. Etude de l'efficacité de l'ivermectine (Mectizan) dans un village d'hyperendémie onchocerquienne de la République Centrafricaine. (In French). Med Afr Noire 1993;40:22-6.
17. Centre Français sur la Population et le Développement. La démographie de 30 états d'Afrique et de l'Océan Indien. (In French). Paris, France: CEPED, 1994.
18. Dumas M, Grunitzky K, Deniau M, et al. Epidemiological study of neuro-cysticercosis in northern Togo (West Africa). Acta Leidensia 1989;57:191-6.
19. Akogun OB, Onwuliri OE. Hyperendemic onchocerciasis in the Tabara River Valley of Gongola State (Old Adamawa Province), Nigeria. Ann Parasitol Hum Comp 1991;66:22-6.
20. Boussinesq M, Chippaux JP, Ernould JC, et al. Effect of repeated treatments with ivermectine on the incidence of onchocerciasis in Northern Cameroon. Am J Trop Med Hyg 1995;53:63-7.
21. Chippaux JP, Boussinesq M, Prod'hom J. Apport de l'ivermectine dans le contrôle de l'onchocercose. (In French). Cah Santé 1995;5:149-58.
22. Diallo S, Lariviere M, Diallo JS, et al. Etude comparative en double aveugle de la tolérance et de l'efficacité de l'ivermectine (MK 933) et du citrate de diéthylcarbamazine (DEC.C) dans le traitement de l'onchocercose humaine. (In French). Med Afr Noire 1985;32:417-37.
23. Bain O, Vuong Ngoc P, Petit G, et al. Différences dans la localisation des microfilaires d'*Onchocerca volvulus* en savane et en forêt : implications cliniques éventuelles. (In French). Ann Parasitol Hum Comp 1986;61:125-6.
24. Ufomadu GO, Eno ROA, Akoh JI, et al. Evaluation of skin biopsies from different body regions of onchocerciasis patients in Central Nigeria. Acta Trop 1988;45:257-61.
25. Jenkins DC. Ivermectin in the treatment of filarial and other nematode diseases of man. Trop Dis Bull 1990;87:R1-9.
26. Preux PM, Melaku Z, Druet-Cabanac M, et al. Cysticercosis and neurocysticercosis in Africa: current status. Neurol Inf Epidemiol 1996;1:63-8.
27. Julvez J, Magnaval JF, Meynard D, et al. Séro-épidémiologie de la toxoplasmose à Niamey, Niger. (In French). Med Trop 1996;56:48-50.
28. Kilian AH. Onchocerciasis and epilepsy. Lancet 1994; 343:983.
29. Kivits M. Quatre cas d'encéphalite mortelle avec invasion d liquide céphalo-rachidien par *Microfilaria loa*. (In French Ann Soc Belge Med Trop 1952;32:235-42.
30. Dumas M, Leger JM, Pestre-Alexandre M. Manifestations neurologiques et psychiatriques des parasitoses. Congrès de Psychiatrie et de Neurologie de Langue Française, session LXXXIV, June 23-27, 1986, Le Mans, France, 1986.
31. Mazzotti L. Presencia de microfilarias de *Onchocerca volvulus* en el liquido cefaloraquideo de enfermos tratados con hetrazam. (In Spanish). Rev Inst Solubr Enferm Trop Mex 1959;19:1-5.

Papers

Mortality associated with HIV-1 infection over five years in a rural Ugandan population: cohort study

Andrew J Nunn, Daan W Mulder, Anatoli Kamali, Anthony Ruberantwari, Jane-Frances Kengeya-Kayondo, Jimmy Whitworth

Abstract

Objective: To assess the impact of HIV-1 infection on mortality over five years in a rural Ugandan population.

Design: Longitudinal cohort study followed up annually by a house to house census and medical survey.

Setting: Rural population in south west Uganda.

Subjects: About 10 000 people from 15 villages who were enrolled in 1989-90 or later.

Main outcome measures: Number of deaths from all causes, death rates, mortality fraction attributable to HIV-1 infection.

Results: Of 9777 people resident in the study area in 1989-90, 8833 (90%) had an unambiguous result on testing for HIV-1 antibody; throughout the period of follow up adult seroprevalence was about 8%. During 35 083 person years of follow up, 459 deaths occurred, 273 in seronegative subjects and 186 in seropositive subjects, corresponding to standardised death rates of 8.1 and 129.3 per 1000 person years. Standardised death rates for adults were 10.4 (95% confidence interval 9.0 to 11.8) and 114.0 (93.2 to 1 per 1000 person years respectively. The mortality fraction attributable to HIV-1 infection was 41% for adults and was in excess of 70% for men aged 25-44 and women aged 20-44 years. Median survival from time of enrolment was less than three years in subjects aged 55 years or more who were infected with HIV-1. Life expectancy from birth in the total population resident at any time was estimated to be 42.5 years (41.4 years in men; 43.5 years in women), which compares with 58.3 years (56.5 years in men; 60.5 years in women) in people known to be seronegative.

Conclusions: These data confirm that in a rural African population HIV-1 infection is associated with high death rates and a substantial reduction in life expectancy.

Introduction

Since the beginning of the HIV-1 pandemic 16 million people are estimated to have become infected in Africa, most of them in sub-Saharan countries.[1] Data on mortality associated with HIV in these countries remain comparatively scarce, and most of the published studies have been in selected hospital or urban populations,[2-6] even though most people in sub-Saharan countries live in rural communities.

In 1994 we published the results of two years of follow up of a rural population cohort in southwest Uganda.[7] We showed the serious impact of HIV-1 infection on this population, in which about 8% of adults are positive for HIV-1. In this paper we report the data for five years of follow up.

Subjects and methods

The area of study is a rural subcounty of Masaka district in southwest Uganda situated about 32 km from Masaka town and 16 km from the trans-African highway. A cluster of 15 villages with a population of about 10 000 was selected for study. The inhabitants are mainly peasant farmers who grow bananas as a subsistence crop and cultivate coffee for sale. The predominant tribal group, the Baganda, constitute about 70% of the population. Substantial numbers of immigrants from Rwanda settled in the area over 20 years ago; more recently some of them have begun to move back to Rwanda.

Late in 1989 the study villages were mapped and an adult member of the household, preferably the head, was asked how many people were in the household. This census included those who had been resident in the household for three months or more, those who had been resident for less than three months but stated that they intended to stay in the area, and those who were regarded as residents but were temporarily living elsewhere—for example, children at boarding school. A socioeconomic questionnaire was also administered.

Within four weeks of these interviews a medical team visited each household. All residents were invited to participate in a survey, which included a brief medical history, a physical examination, and the collection of a blood sample. Absentees and those refusing were revisited to encourage them to participate.

Blood specimens were transported every week to the laboratory of the Uganda Virus Research Institute in Entebbe, where they were tested for antibodies to HIV-1. All serum samples were tested using two enzyme immunoassay systems, Recombigen HIV-1 EIA (Cambridge Biotech, Worcester, MA) and Well-

Medical Research Programme on AIDS in Uganda, Uganda Virus Research Institute, Entebbe
Andrew J Nunn, *senior statistician*
Anatoli Kamali, *epidemiologist*
Anthony Ruberantwari, *statistician*
Jane-Frances Kengeya-Kayondo, *epidemiologist*
Jimmy Whitworth, *head of MRC AIDS programme in Uganda*

Institute of Social Medicine, Academic Medical Centre, University of Amsterdam, 1105 AZ Amsterdam, Netherlands
Daan W Mulder, *former head of MRC AIDS programme in Uganda*

Correspondence to: Mr A J Nunn, MRC HIV Clinical Trials Centre, University College London Medical School, London WC1E 6AU
ajn@mrc.ucl.ac.uk

BMJ 1997;315:767–71

cozyme HIV-1 Recombinant (Wellcome Diagnostics, Dartford, England), western blotting using Novopath HIV Immunoblot (Bio-Rad Laboratories, Watford, England) being used when indicated.[8 9] None of the field workers were aware of the HIV status of study participants. Trained counsellors made results available to all those who requested them.[10]

Every year from 1990 to 1995, the second to sixth rounds of the survey, the census team returned to each household to ascertain the vital status of all those who were resident at the previous survey and to enumerate those who had joined the household through birth or migration. As in the first round of the survey, the medical team collected a blood sample from all those willing to provide one. After the fourth round blood samples were not taken from children (those aged < 13 years).

Monthly birth and death registration was introduced from the beginning of the third round of the survey to supplement data obtained from the annual surveys. An additional question was asked at the time of the census about the vital status of all those who had left the area in the previous 12 months because it is not unusual for seriously ill people to return to their natal home to die.

Statistical methods

Person years of observation were calculated from the time people were enrolled (the date of their first seropositive or seronegative specimen) until the date of the sixth round of the survey for those known to be alive; the date of death was recorded for those known to be dead, and the date of leaving was recorded for those who had left the study area. Those who seroconverted were counted as seronegative until the midpoint between the last known seronegative specimen and first seropositive specimen; thereafter they were counted as seropositive.

Standardised mortality rates were calculated by the direct method with the total population as the standard. Age adjusted mortality rate ratios were calculated by Poisson regression methods. Time to death analyses were performed by using Kaplan-Meier plots and log rank tests.

In the estimates of life expectancy one year intervals were used up to the age of 5 years and five year intervals thereafter, with those aged 75 years or more being combined into one group. In each instance the risk of dying during the first year of life was based on all births reported from the third round of the survey onwards, either through monthly birth and death registration or at the annual census. This is likely to overestimate the risk of dying for seronegative subjects, but inclusion of only those who were known to be seronegative underestimates the risk because the serological state of most of the children dying in the first few months of life was not known.

Results

Of the 1981 households in the study area, 1806 agreed to participate at the initial survey. The total population of these households was 9777 people, of whom 8833 gave a blood sample for HIV-1 testing (either during the initial survey or subsequently) and had at least one unambiguous result. An additional 303 people gave a serum specimen but could not be classified serologically; they were mainly young children whose sample yielded insufficient serum for testing. At the first survey 377 out of 7802 people (4.8%) were HIV-1 positive, 343 (8.2%) of 4172 adults; in all, 412 (4.8%) of 8[illegible] people who were resident at the first survey were fou[illegible] to be positive for HIV-1 at some time, 388 (8.3%) of 4685 adults.

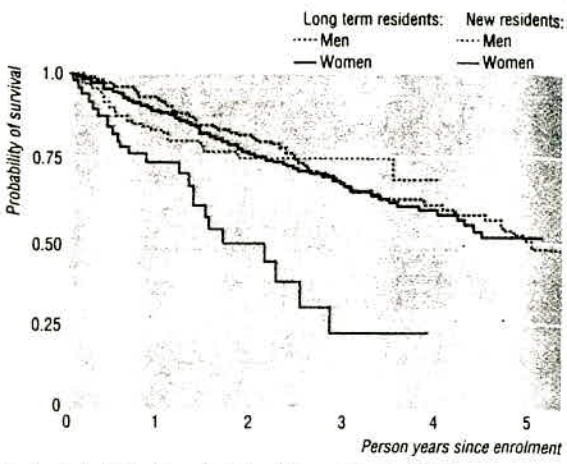


Fig 1 Probability of survival of adults positive for HIV by sex and length of residence

During the five subsequent years a further 5130 people joined the population; 4475 of them were eligible to give a blood sample, and serological status was obtained for 3199 of them.

Of the 9777 people in the first round of the survey, 549 (5.6%) had died by the sixth round and 3090 (31.6%) had left the area and not returned.

A comparison of death rates among seropositive adults resident at the first round of the survey and those joining subsequently showed substantial differences in survival rates (fig 1). Age and sex standardised death rates for those joining were significantly higher than those for members of the population in the first round (hazard ratio after correction for age 1.70 (95% confidence interval 1.24 to 2.32)). This difference was almost entirely on account of substantially higher death rates in the men who joined (age adjusted hazard ratio 3.31; 2.11 to 5.19); the difference in rates betw[illegible] female seropositive residents and those who joined w[illegible] comparatively small (hazard ratio 1.11 (0.73 to 1.71)). Seropositive men who joined the study area seemed to be often in an advanced stage of their disease; hence many of them died soon after their arrival.

Mortality in first round in residents with known HIV status

In the main analysis we compared death rates between seronegative and seropositive subjects in the cohort resident during the first round of the survey. A total of 459 deaths was reported in 35 083 person years of observation (table 1); 273 were seronegative subjects and 186 seropositive, corresponding to 8.1 and 129.3 per 1000 person years. Standardised death rates among seronegative and seropositive adults were 10.4 (9.0 to 11.8) and 114.0 (93.2 to 134.8) per 1000 person years respectively.

The excess mortality due to HIV-1 infection in adults (mortality in total adult population minus mortality in people negative for HIV-1) was 7.4 (7.5

Age (years)	HIV positive			HIV negative			Total		
	No of person years	No of deaths	Rate*	No of person years	No of deaths	Rate*	No of person years	No of deaths	Rate*
0 -	3 161	39	12.3	33	13	397	3 194	52	16.3
5 -	10 992	28	2.5	55	5	90	11 047	33	3.0
13 -	6 359	13	2.0	76	2	26	6 435	15	2.3
20 -	2 086	11	5.3	252	21	83	2 338	32	13.7
25 -	3 333	16	4.8	541	71	131	3 875	87	22.5
35 -	2 261	10	4.4	252	30	119	2 513	40	15.9
45 -	2 061	23	11.2	126	17	135	2 187	40	18.3
≥55	3 391	133	39.2	104	27	260	3 495	160	45.8
Total	33 645	273	8.1	1 439	186	129	35 083	459	13.1
Adult (≥13)	19 492	206	10.6	1 351	168	124	20 842	374	17.9

*Number of deaths per 1000 person years.

when age standardised (6.4 to 8.7)) (table 2). The excess was highest (17.7 per 1000) in those aged 25-34. The mortality fraction attributable to HIV-1 infection was 41% for adults (those aged 13 and over) and 69% for those aged 13-44.

Standardised death rates in seropositive men and women were similar, being 104.0 and 118.1 per 1000 son years; excess mortality and attributable tality fractions varied by sex and age (table 2), reflecting to a large extent differential rates of seropositivity. Thus, in those aged 20-24 the attributable mortality fraction was very high for women (76%) but much lower for men (9%). For all women combined the attributable mortality fraction was 42% compared with 40% in men.

The HIV-1 specific age adjusted mortality rate ratios (comparing HIV positive subjects with HIV negative subjects) were 11.9 (8.3 to 15.9) for men and 13.9 (10.3 to 18.8) for women; the corresponding ratios for those aged 13-44 years were 16.4 (9.9 to 27.4) and 28.0 (17.2 to 45.6)

Among adults who were positive for HIV-1 there was no difference in the rate of survival of men and women (fig 1) ($\chi^2 = 0.29$, 1 df, $P = 0.6$, log rank test). When the data were analysed by age group (combining those aged 13-19 and 20-24) there were no differences between the four groups aged under 55 years of age ($\chi^2 = 6.15$, 3 df, $P = 0.14$) (fig 2). However, those aged 55 ars or more died considerably more rapidly than those aged 13-54 years ($\chi^2 = 16.28$, 1 df, $P = 0.0001$).

Two years after enrolment the probabilities of survival for those aged 13-54 and 55 or more were 81.3% (77.4% to 84.7%) and 53.6% (37.7% to 67.2%) respectively. By five years the survival probabilities were 53.1% (45.4% to 60.2%) and 28.9% (14.1% to 45.6%) respectively. The hazard ratio comparing survival rates in these two age groups was 2.30 (1.52 to 3.47).

The crude death rate for the total resident population regardless of HIV positivity was 14.6 per 1000 person years overall and 19.3 per 1000 person years in adults. Standardised rates in adults without a defined serostatus were considerably higher than in those with: 39.0 and 17.9 respectively. The rate in the former group was particularly high during the first year of follow up (87.9 per 1000 person years) but declined thereafter to 26.6 per 1000 person years for the remainder of the follow up period.

Table 2 Excess mortality and mortality fraction attributable to HIV-1 infection in adults by sex

Age group (years)	Excess mortality per 1000			Mortality fraction attributable to HIV-1 (%)		
	Males	Females	Total	Males	Females	Total
13-19	0.0	0.6	0.3	0.0	20.5	12.3
20-24	0.5	16.6	8.4	9.0	76.2	61.5
25-34	19.7	15.8	17.7	78.5	78.6	78.6
35-44	15.9	7.8	11.5	72.1	71.4	72.2
45-54	7.7	6.5	7.1	33.7	45.0	39.0
≥55	8.0	5.1	6.6	18.1	10.7	14.3
Total	7.3	7.5	7.4	40.4	41.8	41.1
13-44	7.3	8.5	7.9	66.2	71.5	69.0

Life expectancy

Life expectancy in people resident in the study area at any time between the first and sixth rounds was estimated to be 42.5 years (41.4 for men and 43.5 for women). For those known to be seronegative, life expectancy was estimated as 58.6 years (56.5 years for men and 60.5 years for women); these estimates were based on 56 000 and 47 000 person years of observation respectively.

The probability of a 15 year old dying before reaching the age of 60 is 0.61 in the total population and 0.24 in people known to be seronegative: in men the probabilities are 0.66 and 0.30 and in women 0.57 and 0.19.

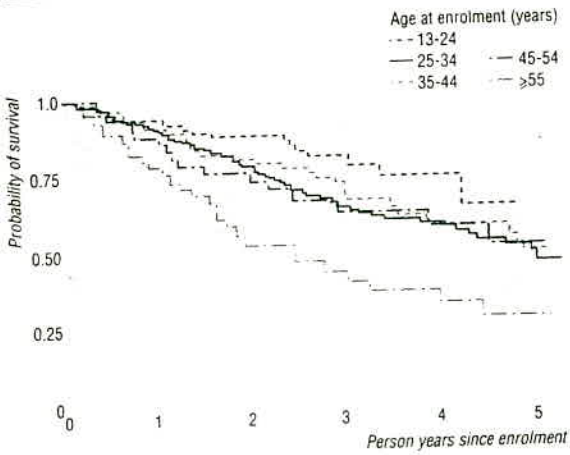


Fig 2 Probability of survival of adults positive for HIV-1 after five years of follow up

Discussion

In 1994 we reported death rates associated with HIV-1 infection in a rural Ugandan population after a follow up of two years with about 16 000 person years of observation.[7] This report extends follow up to five years and is based on more than twice as many person years of observation.

We have confirmed our earlier findings of the impact of the HIV-1 epidemic on this rural population, whose underlying stable seroprevalence is about 5% and about 8% among adults. The mortality attributable to HIV-1 infection is estimated to be 41% in adults and over 70% in women aged 20-44 and men aged 25-44, which reflects the differences in age specific seroprevalence between the two sexes.

Possible biases

Possible biases that should be considered when interpreting these results include selective enrolment and selective migration out of the area. Of the 9777 residents in the first round (in whom these estimates are based) 90% were enrolled and included in the analysis because they agreed to have blood samples taken and had an unambiguous result on testing for HIV-1 antibody. The age standardised mortality among the adults in the census who did not have blood samples taken was 39.0 per 1000 person years, considerably higher than the combined rates for seropositive and seronegative subjects for each age group. This could be due to a higher seropositivity rate in those who did not comply, but it may also reflect a high proportion of fatally ill people because 32 of the 62 adults who died in this group died within the first year of follow up.

Seriously ill patients often return to the village of their birth to die, and thus migration out of this closed cohort is likely to have resulted in an underestimate of the number of deaths and therefore of death rates. To minimise the number of deaths not ascertained, we inquired at each annual census from the third round onwards about whether those who had left the area in the previous 12 months were known to have died. The monthly death registration, introduced at the same time, did not identify any deaths that were not subsequently ascertained at the annual census.

We cannot rule out the possibility of errors in the ascertainment of HIV-1 antibody in the laboratory, but most participants had their blood assessed on more than one occasion. The HIV testing algorithm used had a sensitivity and specificity close to 100%.[6]

Key messages

- Comparatively few data exist on mortality associated with HIV-1 in sub-Saharan Africa
- Adults positive for HIV-1 in a rural Ugandan population with a prevalence of infection of 8% were more than 10 times more likely to die over a 5 year period than those negative for HIV-1
- Over 40% of all deaths in adults were attributable to HIV-1 infection, the percentage in young adults aged 25-44 being in excess of 70%
- Life expectancy is estimated to have declined from 58.6 to 42.5 years as a consequence of the AIDS epidemic

Mortality associated with HIV

The extremely high mortality risk ratios associated with HIV-1 infection provide strong evidence that HIV-1 is the cause of substantial excess mortality.[11] The age adjusted risk ratio for adults in this study, 13.2 (10.6 to 16.4), compares with recent reports on relative risks of 9.5 (6.0 to 14.9) in a general population in the neighbouring Rakai district of Uganda[12]; of 12.9 (5.4 to 30.7) in an occupational cohort in Mwanza, Tanzania[6]; and of 13.3 (10.0 to 17.2) in people with haemophilia in Britain.[13] The relative risks observed in different populations using different study methods are remarkably similar, thus adding weight to the causal association between HIV and excess mortality.

It has been suggested that immunosuppressive foreign proteins contaminating commercial factor VIII and treatment with zidovudine may be causes of AIDS.[14,15] Sabin et al, comparing HIV negative and HIV positive men with haemophilia A, recently rejected this hypothesis,[16] but the debate continues.[17] None of our subjects had received factor VIII or zidovudine; nor was there any evidence of misuse of injected drugs in the population.

The proportion of deaths attributable to HIV this rural population was 42% among women and 40% among men. The proportions were even higher in those aged 13-44, 72% for women and 66% for men, peaking at close to 80% in both sexes in those aged 25-34. Similar mortality fractions attributable to HIV have been reported in a cohort of women of childbearing age in Rwanda[5]; in an occupational cohort in Mwanza, Tanzania[6]; and in the rural stratum of a population study in Rakai District, Uganda.[12] The profound impact of the HIV-1 epidemic on mortality in rural Uganda is also shown by estimates of life expectancy at birth, which is now only 42.5 years compared with 58.6 years in those who are uninfected. The probability of a 15 year old surviving to the age of 60 (0.39) is about half that of a seronegative person of the same age (0.76).

Survival

We previously noted the rapid progression from asymptomatic infection or mild disease to death[7]; of half of the patients for whom data were available l one or more major symptoms of AIDS at the medical assessment one year or less before death. The five year follow up data show that infected people of 55 years or more progress to death much more rapidly than younger people (estimated median survival <3 years v >5 years). Increased rates of progression with age have been observed in several industrialised countries.[16-21]

No clear trend in survival rate by age group emerged from the analysis of those aged 13-54 who were HIV antibody positive. Only as a larger cohort of those who have seroconverted is followed up for a longer time will possible differences in survival by age in this group become apparent.

Results from studies of survival after infection with HIV-1 among haemophilic and homosexual populations in North America and Europe suggest a median survival time from infection to death of 9-11 years.[22-24] Comparatively little is known about survival rates in developing countries. The results of some early studies suggested rates similar to those in industrialised countries,[5,25-27] but a recent study among prostitutes in

Nairobi also documented a much faster rate of progression, corresponding to a median survival from infection to AIDS of 4.4 years.[28] We found similar rates for men and women; this again is consistent with reports from industrialised countries.[29-32]

A recent study describing progression to AIDS and survival after the diagnosis of AIDS in Africans living in London found that their survival was more similar to that of patients born in industrialised countries than to that of patients living in Africa.[33] However, the cohort studied was retrospective, the time of follow up was short, and a fifth of patients were lost to follow up for more than a year, including some who had returned to Africa. Although such studies are useful, large scale prospective studies are needed to understand better the survival experience of people of different ethnic origins living in different countries.

Funding: This study was supported by the Medical Research Council and Overseas Development Administration of the United Kingdom.Conflict of interest: None.

inn TC. Global burden of the HIV pandemic. *Lancet* 1996;348:99-106.
nn JM, Francis H, Quinn T, Nzilambi N, Bosenge N, Bila K, et al. Surveillance for AIDS in a central African city—Kinshasa, Zaire. *JAMA* 1986;255:3255-9.
3 Nelson AM, Hassig SE, Kayembe M, Okonda L, Mulanga K, Brown C, et al. HIV-1 seropositivity and mortality at University Hospital, Kinshasa, Zaire, 1987. *AIDS* 1991;5:583-6.
4 De Cock KM, Barrere B, Diaby L, Lafontaine M-F, Gnaore E, Porter A, et al. AIDS—the leading cause of death in the west African city of Abidjan, Ivory Coast. *Science* 1990;249:793-6.
5 Lindan CP, Allen S, Serufilia A, Lifson AR, Van de Perre P, Chen-Rundle A, et al. Predictors of mortality among HIV-infected women in Kigali, Rwanda. *Ann Int Med* 1992;116:320-8.
6 Borgdorff MW, Barongo L, Klokke AH, Newell JN, Senkoro KP, Velma JP, et al. HIV-1 incidence and HIV-1 associated mortality in a cohort of urban factory workers in Tanzania. *Genitourin Med* 1995;71:212-5.
7 Mulder DW, Nunn AJ, Kamali A, Nakiyingi J, Wagner H-U, Kengeya-Kayondo JF. Two year HIV-1 associated mortality in a Ugandan rural population. *Lancet* 1994;343:1021-3.
8 Nunn AJ, Biryahwaho B, Downing RG, van der Groen G, Ojwiya A, Mulder DW. Algorithms for detecting antibodies to HIV-1: results from a rural Ugandan cohort. *AIDS* 1993;7:1057-61.
9 Nunn AJ, Biryahwaho B, Downing RG, Ojwiya A, Mulder DW. Computer-assisted quality assurance in a HIV serology laboratory. *Meth Inform Med* 1994;33:170-3.
10 Seeley JA, Wagner H-U, Mulemwa J, Kengeya-Kayondo JF, Mulder DW. The development of a community based HIV/AIDS counselling service in a rural area in Uganda. *AIDS Care* 1991;3:207-17.
11 Dondero TJ, Curran JW. Excess deaths in Africa from HIV: confirmed and quantified. *Lancet* 1994;343:989-90.
ankambo NK, Wawer MJ, Gray RH, Serwadda D, Li C, Stallings RY. ographic impact of HIV infection in rural Rakai district, Uganda: sults of a population based cohort study. *AIDS* 1994;8:1707-13.
13 Darby SC, Ewart DW, Giangrande PLF, Dolin LF, Spooner PJ, Rizza RJD. Mortality before and after HIV infection in the complete UK population of haemophiliacs. *Nature* 1995;377:79-82.
14 Duesberg P. Foreign protein-mediated immunodeficiency in haemophiliacs with and without HIV. *Genetica* 1995;95:51-70.
15 Duesberg P. Is HIV the cause of AIDS? *Lancet* 1995;346:1371-2.
16 Sabin CA, Pasi KJ, Phillips AN, Lilley P, Bofill M, Lee CA. Comparison of immuodeficiency and AIDS defining conditions in HIV negative and HIV positive men with haemophilia A. *BMJ* 1996;312:207-10.
17 Duesberg P. Commentary: non-HIV hypotheses must be studied more carefully. *BMJ* 1996;312:210-1.
18 Carre N, Deveau C, Belanger F, Boufassa F, Persoz A, Jadand C, et al. Effect of age and exposure group on the onset of AIDS in heterosexual and homosexual HIV-infected patients. *AIDS* 1994;8:797-802.
19 Rosenberg PS, Goedert JJ, Biggar RJ for the multicenter hemophilia cohort study and the international registry of seroconverters. Effect of age at seroconversion on the natural incubation distribution. *AIDS* 1994;8:803-10.
20 Blatt SP, McCarthy WF, Bucko-Krasnicka B, Melcher GP, Boswell RN, Dolan MJ, et al. Multivariate models for predicting progression to AIDS and survival in human immunodeficiency virus-infected persons. *J Inf Dis* 1995;171:837-44.
21 Martin JN, Colford JM Jr, Ngo L, Tager IB. Effect of older age on survival in human immunodeficiency virus (HIV). *Am J Epidemiol* 1995;142:1221-30.
22 Hendriks JCM, Medley GF, Van Griensven GJP, Coutinho R, Heisterkamp SH, van Druten HAM. Treatment-free incubation period of AIDS in a cohort of homosexual men. *AIDS* 1993;7:231-9.
23 Veugelers PJ, Page KA, Tindall B, Schecter MT, Moss AR, Winkelstein WW Jr, et al. Determinants of HIV disease progression among homosexual men registered in the tricontinental seroconverter study. *Am J Epidem* 1994;140:747-58.
24 Schechter MT, Le Nhu, Craib KJP, Le TN, O'Shaughnessy MV, Montaner JSG. Use of the Markov model to estimate waiting times in a modified WHO staging system for HIV infection. *J AIDS* 1995;8:474-9.
25 N'galy B, Ryder RW, Kapita B, Mwandagalirwa K, Colebunders RL, Francis H, et al. Human immunodeficiency virus infection among employees in an African hospital. *N Engl J Med* 1988;319:1123-7.
26 Hira SK, Ngandu N, Wadhawan D, Nkowne B, Baboo KS, Macuacua R, et al. Clinical and epidemiological features of HIV infection at a referral clinic in Zambia. *J AIDS* 1990;3:87-91.
27 Bulterys M, Nzabihimana E, Chao A, Bugingo G, Muskera J, Saah A, et al. Long-term survival among HIV-1 infected prostitutes. *AIDS* 1993;7:1269.
28 Anzala OA, Nico JD, Nagelkerke NJD, Bwayo JJ, Holton D, Moses S, et al. Rapid progression to disease in African sex workers with human immunodeficiency virus type 1 infection. *J Infect Dis* 1995;171:686-9.
29 Melnick SL, Sherer R, Louis TA, Hillman D, Rodriguez EM, Lackman C, et al. Survival and disease progression according to gender of patients with HIV infection. *JAMA* 1994;272:1915-21.
30 Von Overbeck J, Egger M, Davey Smith G, Schoep M, Ledergerber B, Furrer H, et al. Survival in HIV infection: do sex and category of transmission matter? *AIDS* 1994;8:1307-13.
31 Phillips AN, Antunes F, Stergious G, Ranki A, Jensen GF, Bentwich Z, et al. A sex comparison of rates of new AIDS-defining disease and death in 2554 AIDS cases. *AIDS* 1994;8:831-5.
32 Lepri AC, Pezzotti P, Dorrucci M, Phillips AN, Rezza G. HIV disease progression in 854 women and men infected through injecting drug use and heterosexual sex and followed for up to nine years from seroconversion. *BMJ* 1994;309:1537-42.
33 Low N, Paine K, Clark R, Mahalingham M, Pozniak A. AIDS survival and progression in black Africans living in south London, 1986-1994. *Genitourin Med* 1996;72:12-6.

(Accepted 19 May 1997)

Any questions

Physical activity is best for back pain

Several of my patients who suffer with low back pain and stiffness tell me that swimming or static cycling temporarily ease their symptoms. What is the explanation for this?

The traditional recommendation for people with back problems is periods of bed rest until they resolve. We now know that advice is wrong, and indeed for many patients rest aggravates the back problem and perpetuates disability. Careful controlled studies have shown that limited, if any, periods of bed rest are best and physical activity as soon as possible is recommended.[1]

Exercise may take many forms. The precise nature of the exercise seems relatively unimportant, but for most patients any form within reason and which does not cause major aggravation seems helpful. The mechanisms by which exercise provides relief are unclear. Venous congestion and oedema around the nerve roots within the vertebral canal and intervertebral foramen develop in relation to mechanical damage in the back.[2] Physical exercise is likely to temporarily relieve this in much the same way as exercise is good for people with varicose veins and oedema of the feet.

Muscular activity will cause an afferent discharge, which may inhibit the transmission of the pain signal in the spinal cord in a similar way as rubbing the site of an injury will relieve local pain. Movement of the spine likewise will lead to afferent impulses, which may relieve symptoms in a similar fashion. This may be analogous to the relief of symptoms on exercising a stiff and painful peripheral joint. Consideration of this type of mechanism led to the use of electrical stimulation for the relief of chronic pain.

Malcolm I V Jayson, *professor of rheumatology, Manchester*

1 Clinical Standards Advisory Group. *Back pain.* London: HMSO, 1994.
2 Hoyland JA, Freemount AJ, Jayson MIV. Intervertebral foramen venous obstruction. *Spine* 1989;14:558-68.

American Journal of Epidemiology

Vol. 149, No. 6
Printed in U.S.A.

Onchocerciasis and Epilepsy: A Matched Case-Control Study in the Central African Republic

Michel Druet-Cabanac,[1] Pierre-Marie Preux,[1,2] Bernard Bouteille,[1,3] Patricia Bernet-Bernady,[1] Jean Dunand,[4] Adrian Hopkins,[5] Georges Yaya,[4] André Tabo,[4] Chiara Sartoris,[6] Waruingi Macharia,[1] and Michel Dumas[1]

The occurrence of epileptic seizures during onchocercal infestation has been suspected. Epidemiologic studies are necessary to confirm the relation between onchocerciasis and epilepsy. A matched case-control study was conducted in dispensaries of three northwestern towns of the Central African Republic. Each epileptic case was matched against two nonepileptic controls on the six criteria of sex, age (±5 years), residence, treatment with ivermectin, date of last ivermectin dose, and the number of ivermectin doses. Onchocerciasis was defined as at least one microfilaria observed in iliac crest skin snip biopsy. A total of 561 subjects (187 cases and 374 controls) were included in the study. Of the epileptics, 39.6% had onchocerciasis, as did 35.8% of the controls. The mean dermal microfilarial load was 26 microfilariae per mg of skin (standard deviation, 42) in the epileptics and 24 microfilariae per mg of skin (standard deviation, 48) in the controls. This matched case-control study found some relation (odds ratio = 1.21, 95% confidence interval 0.81-1.80), although it was nonstatistically significant. *Am J Epidemiol* 1999;149:565–70.

case-control studies; epilepsy; microfilarial load; onchocerciasis

Epileptic seizures are frequently mentioned in relation to many parasitic infestations, although it has been difficult to demonstrate any clear cause-effect relation. Seizures are frequent in neurocysticercosis and may occur during cerebral malaria.

They are not described in the treatises of tropical medicine as occurring during onchocercal infestation, the filariasis causing river blindness. As early as 1919, Roblès (1) reported the case of an epileptic patient presenting with an onchocercal nodule that had perforated the cranial vault. The possibility of a relation between onchocerciasis and epilepsy was highlighted by Puyelo and Holstein (2) and then by Rwiza (3) and Jilek-Aall (4), who had noted that the prevalences of epilepsy, mental illness, and growth retardation were higher in onchocercal hyperendemic zones. Several cross-sectional studies (5-8) and one case-control study (9) were conducted to confirm this relation. The results were, however, contradictory, and the necessity for further studies was felt among the workers.

The objective of this study, which was conducted in the Central African Republic, was to establish whether there exists an eventual link between *Onchocerca volvulus* infestation and epilepsy.

MATERIALS AND METHODS

The study was performed in March 1996, in the savannah woodland region in the northwest Central African Republic. This region is classified by the National Programme for the Fight Against Onchocerciasis and Blindness (NPFOB) as a hyperendemic zone for onchocerciasis (about 70 percent of the people are infested). A matched case-control design was used. The cases were epileptics subjects, and the controls were nonepileptics. The exposure factor studied was onchocerciasis.

Clinical and paraclinical examinations were carried out on the epileptics and the controls in the dispensaries of three towns (Bossangoa, which is situated in the valley of Ouham River, and Paoua and Ngaoundaye, situated in the valley of Ouham-Pendé

Received for publication January 28, 1998, and accepted for publication July 9, 1998.

Abbreviations: ESB, excisional skin biopsies; HIV, human immunodeficiency virus; MFL, mean dermal microfilarial load; NPFOB, National Programme for the Fight Against Onchocerciasis and Blindness.

[1] Institute of Epidemiology and Tropical Neurology, Faculty of Medicine, Limoges, France.
[2] Department of Biostatistics and Medical Informatic, Faculty of Medicine, Limoges, France.
[3] Department of Parasitology, Faculty of Medicine, Limoges, France.
[4] Faculty of Medicine, Bangui, Central African Republic.
[5] National Programme for the Fight Against Onchocerciasis and Blindness, Bossangoa, Central African Republic.
[6] Deutsche Gesellschaft für Technische Zusammenarbeit, Bossangoa, Central African Republic.
Reprint requests to Dr. Pierre-Marie Preux, Institut d'Epidémiologie neurologique et de Neurologie Tropicale, Faculté de Médecine, 2 rue Dr. Marcland, 87025 Limoges, France.

River). The subjects hailed from 66 villages in the region.

Inclusion

The medical personnel working in these towns had, in the weeks preceding the study, notified the village leaders, the epileptics, and their families about their participation in the study. The study was carried out under the auspices of the Central African Republic's Ministry of Population and Public Health. Informed verbal consent was obtained from each included subject.

The epileptics were defined as subjects age 15 years and above who had had two or more unexplained seizures (10) (with all types of generalized or partial seizures included). The controls were age 15 years and above and did not have a neurologic illness.

Matching criteria

Two nonepileptic controls were matched with one epileptic for the factors that may affect the dermal microfilarial load (MFL). These included age (±5 years), sex, geographic zone of residence, whether or not they had been treated with ivermectin (i.e., treatment used in the annual mass treatment of onchocerciasis), the number of the doses, and the date of the last dose of ivermectin. After the matching was verified, each triplet, thus formed, was examined by a neurologist who verified the clinical inclusion and exclusion criteria.

Neurologic examination

A neurologist, helped by a local interpreter, confirmed the diagnosis after a thorough history taking (using a questionnaire designed by the Institute of Epidemiology and Tropical Neurology, Limoges, France (11), as recommended by the World Health Organization (12), in collaboration with the Pan African Association of Neurological Sciences) and a clear neurologic clinical examination and categorized the seizures according to the classification defined by the Commission on Classification and Terminology of the International League Against Epilepsy (10). Motor and sensory problems; cerebellar, pyramidal, and extrapyramidal syndromes; as well as all cranial nerves palsies were thoroughly evaluated in both the cases and the controls. The presence of abnormal physical signs led to exclusion of a subject. The neurologists looked for the dates of onset of seizure disorder, the date of the last seizure, whether the epilepsy was active, the seizure frequency before and after ivermectin therapy, the antiepileptic drugs used, the positive family history of epilepsy, and a positive personal history that could explain the occurrence of seizures (e.g., fetal or neonatal trauma, prematurity at birth, cranial trauma, meningitis, etc.).

Ophthalmologic examination

All subjects were examined by an ophthalmologist, who measured their visual acuity (using the Snellen chart test) and intraocular pressure (by Schiotz tonometry). Onchocercal-specific and nonspecific ocular lesions (e.g., conjunctivitis, punctate or sclerosing keratitis, uveitis, chorioretinis, and optic atrophy) were checked for by using an ophthalmoscope and a slit lamp.

Dermatologic examination

Cutaneous lesions specific to onchocerciasis (e.g., onchocerca dermatitis, cutaneous atrophy, dyspigmentation, and subcutaneous nodules) were thoroughly searched for by inspection and palpation. These examinations were carried out in both the cases and the controls.

Parasitologic examination

Excisional skin biopsies (ESB) were taken from each iliac crest by using a Walzer sclerotomy skin snipper. A sample of 30 ESBs was weighed, and the mean weight was defined as 1 mg per ESB. Each ESB was left to stand in wells of a microplate containing 100 μl of sterile 0.9 percent sodium chloride. After 4 hours of incubation, the parasitologist performed a physical count of the *O. volvulus* microfilariae under light microscopy to determine the mean dermal MFL for each person. A subject was considered to be infested with onchocerciasis if at least one microfilaria was found in either of the two ESBs. For detection of the adult form of the parasite in the subjects who had nodules, eight nodulectomies were performed in the operating theaters at the dispensaries.

Serologic evaluation

A sample of 10 ml of blood as drawn from each subject by using nonanticoagulated Vacutainer tubes (Becton Dickinson, Franklin Lakes, New Jersey). Each sample was then centrifuged, and the sera were transferred into cryofreezing Nunc tubes (Nunc A/S, Roskilde, Denmark) and immediately put into liquid nitrogen containers. Then they were conserved in dry ice in an isothermal container ready to be flown to Limoges, France. Cysticercosis, human immunodeficiency virus (HIV), and toxoplasmosis serologies were

performed on all of the sera. The screening for cysticercosis was done by enzyme-linked immunosorbent assay method using a crude cysticercal antigen prepared as described by Guerra et al. (13) and Chamouillet et al. (14). The screening for HIV was done using enzyme-linked immunosorbent assay Uniform II EIA Kit (Organon, Fresnes, France). The confirmation of seropositivity to the HIV-1 was done by Western blot (Biotec/Dupont HIV-1 immunoglobulin G Western blot, Ortho Diagnostic Systems, Roissy, France), and that of HIV-2 was performed simultaneously in the New Lav-Blot 2 (Diagnostic Pasteur, Marnes la Coquette, France). The screening for toxoplasmosis was done using the enzyme immunoassay method with a final fluorescent detection (Vidas Toxo Immunoglobulin G, Mérieux, Marcy-l'Etoile, France).

Statistical analysis

Data was analyzed using Epi-Info 5.01 b software (Centers for Disease Control and Prevention, French version, National School of Public Health, Atlanta, Georgia, 1992) and Statview 4.5 software (Abacus Concept, Inc., Berkeley, California). Quantitative variables were given in the form of mean (standard deviation). The frequency comparisons were made by Pearson chi-square or Fisher exact tests. The mean comparisons were performed by using the Student t test, the Mann-Whitney U test, or the Kruskall-Wallis test. The correlations were done by correlation coefficient calculation or by Spearman rank test. Exposure proportions of the cases and the controls were calculated, and the matched odds ratios and their confidence intervals were estimated. Each confidence interval was mated to the risk of 5 percent. The minimum number of subjects necessary for the study (two controls for each case with an error risk α at 5 percent and a power of 80 percent, in a region where the prevalence of onchocerciasis in the controls was estimated at 70 percent) was 176 triplets and, hence, 528 subjects.

RESULTS

Demographic data

A total of 187 triplets were included, representing 187 epilepsy cases and 374 nonepileptic controls, for a total of 561 subjects. Of these, 225 (40.1 percent) were females and 336 (59.9 percent) were males. The mean age was 25.6 years (8.6 standard deviation (SD)).

Onchocerciasis

Among the 561 subjects examined, 208 were positive for onchocerciasis. This corresponded to 37.1 percent of the study population (38.4 percent were males and 35.1 percent were females). The mean age was 26.3 (8.7 SD) years in those who had onchocerciasis and 25.3 (8.5 SD) years in those who did not. The mean dermal MFL in the subjects was 25 (40 SD) microfilariae per mg of skin. Females had an MFL of 31 (51 SD) microfilariae per mg of skin, while males had an MFL of 21 (42 SD) microfilariae per mg of skin. The ivermectin therapeutic cover was 89.7 percent. In the study population, 503 of 561 subjects had taken ivermectin at least once and on average were treated twice. The mean period between the last ivermectin dose and the date of entry into the study was 5 months. This period varied between 15 days and 34 months. The mean dermal MFL was 26 (46 SD) in those already treated with ivermectin microfilariae per mg of skin and 36 (42 SD) microfilariae per mg of skin in the nontreated cases ($p < 0.01$).

Cutaneous nodules were found in 87 of 561 examined subjects, representing 15.5 percent of the study population. More than 75 percent of the nodules were located at the level of the iliac crest, and no nodule was found on the head. Sixty-six percent of the subjects with nodules had onchocerciasis. The proportion of nodule carriers was higher in those who had onchocerciasis ($p < 0.001$). Seven of eight nodules biopsied were onchocercomas. The eighth was a mature lipoma without microfilaria.

Blindness was found in 1.2 percent of the examined subjects. Six of the seven blind persons were affected in both eyes. The causes of blindness were multiple: two subjects had onchocerciasis, one had glaucoma, two had a sclerosing keratitis, and two had anterior uveitis nonspecific for onchocerciasis.

Epilepsy

The mean age of epileptics was 25.0 (7.6 SD) years (23.6 (6.3 SD) years for females and 26.1 (8.2 SD) years for males). The distribution of the epileptic seizures was 95.1 percent generalized seizures, 1.2 percent simple partial seizures, and 3.7 percent complex partial seizures. In this study, 96.3 percent of the subjects had active epilepsy. The mean evolution period for the epilepsy was 9 years (7 SD). The mean period between the last seizures and the first day of the study was 4 months (10 SD); this varied between 0 and 71 months. A positive family history of epilepsy was found in 32.6 percent of the epileptic patients, 7.4 percent had suffered measles with complications, 1.6 percent had suffered from meningitis or encephalitis, and 3.7 percent had difficulties at birth; 10.2 percent of the mothers of the epileptics had some problems during pregnancy. However, 41 percent of the epileptics had no relevant personal or family history that could explain their illness.

Relation between onchocerciasis and epilepsy

There was no significant difference between the cases and the controls on the matching criteria of age, sex, residence, treatment with ivermectin, number of ivermectin doses, and date of last ivermectin dose. Sixty-six percent of 187 triplets were perfectly matched for the six criteria, but all were perfectly matched for at least two matching criteria (sex and residence).

The matched odds ratio was 1.21 (95 percent confidence interval 0.81-1.80). Of the epileptics, 39.6 percent had onchocerciasis, as did 35.8 percent of the controls. The mean dermal MFL was 26 (42 SD) microfilariae per mg of skin in the epileptics and 24 (48 SD) microfilariae per mg of skin in the controls. Thus, there was no significant difference between the epileptics and the controls.

The number of matching criteria did not significantly modify the results of this study, since the matched odds ratio varied from 1.10 for the triplets matched for all six criteria to 1.21 for the triplets matched for at least two criteria. Screening for other factors that could explain epileptic seizures, such as HIV infection, cysticercosis, and toxoplasmosis infestation, did not reveal a significant difference between the cases and the controls (3 percent of the epileptics and 7 percent of the controls tested positive for HIV, 3.7 percent of the epileptics and 2.4 percent of the controls were seropositive for cysticercosis, and 15 percent of the epileptics and 18 percent of the controls were seropositive for toxoplasmosis).

DISCUSSION

Few studies on the prevalence of onchocerciasis in the Central African Republic have been published (15, 16). The NPFOB usually estimates this prevalence before the ivermectin mass treatment campaigns. To our knowledge, no study on epilepsy has been carried out in this country. Onchocerciasis and epilepsy represent two major health problems in the Central African Republic, particularly because of their serious medical, social, cultural, and economic implications. The local medical team believed that the prevalence of epilepsy was high. We studied these two diseases to discover if any causal relation exists between them.

The village leaders and families of the epileptics were notified in the weeks preceding this study. This allowed for sufficient inclusion of the cases and the controls to attain the minimal number of subjects. The mean sample age was 25.6 years (8.6 SD), which was comparable with the mean age of the general population (17). The sex ratio was 1.49. This sex disproportion at inclusion is probably related in part to the fact that women of marriage age are wary of confessing their epilepsy (18). The matching of cases and controls on the different factors that could affect the dermal MFL increases the study validity. The matched factors were as follows: age (as MFL increases with age) (16), sex (with MFL being higher in males than in females) (19, 20), residence (because the MFL is clearly correlated with the distance between the river and the homesteads) (20), treatment with ivermectin (21), and the dose and timing of treatment with ivermectin (22).

Onchocerciasis was defined in the cases and controls upon a positive ESB, and the mean dermal MFL was estimated. The latter method was used here for the first time to determine the link between epilepsy and onchocerciasis. The screening of onchocerciasis was performed by two ESBs taken at the two iliac crest (23) sites where the sensitivity of the technique for searching for microfilaria was 95 percent (24).

The prevalence of onchocerciasis of 37.1 percent in the controls was much lower than that found previ by NPFOB. This is probably linked to the excell therapeutic cover, since 89.7 percent of subjects studied had at least had one cycle of ivermectin treatment with two doses of ivermectin on average. Testa et al. (16) and Diallo et al. (22) had shown during the ivermectin therapeutic trials that the MFL was reduced by more than 80 percent during the days that immediately followed a dose of ivermectin but remained detectable in some ESBs. Ivermectin causes intrauterine degeneration in the adult worms and temporary sequestration of unborn microfilariae (25). In the subsequent months, the adult worms resume microfilarial production, and the MFL increase again. In our study, the mean period of the last ivermectin dose was 5.5 months. One would suppose, therefore, that the majority of the people with treated onchocerciasis were positive during the screening. However, it was possible that some subjects classified as negative for onchocerciasis actually had onchocerciasis. In this case, the probability of fir false negatives should not have been different between the cases and the controls. Only in a single case did the ophthalmologic examination reveal microfilaria, which were few in number, in the anterior chamber of the eye when the iliac crest ESB was negative.

Cysticercosis, HIV, and toxoplasmosis infection were also screened for, but did not seem to explain the epilepsy cases in this study. In the epileptic subjects, the HIV prevalence was 6 percent, which is remarkably high for a rural area. The prevalence of cysticercosis was 2.9 percent, and toxoplasmosis was 17 percent. This results did not differ from the levels previously found in central Africa (26, 27).

This study had the aim of determining whether there is a relation between *O. volvulus* infestation and

Strategies of analysis

Review classical techniques

Understand rationale for moving to regression models

Identifying potential variables to be included in model

Strategies of model building

Stages of data analysis

Preliminary analysis:

- Data cleaning
- Classification of variables
- Data reduction

Crude analysis:

- Tables
- Crude measure of effect

Bivariable analyses (adjust for 1 confounder at a time):

- Stratified tables
- Assess for effect modification and confounding
- Adjusted measure of effect (if no interaction)

Multivariable analysis (adjust for >1 confounders)

- "Fully" adjusted measure of effect

Question: In Mwanza dataset, examine effect of number of partners in the past year on odds of HIV infection adjusting for potential confounders.

Cases: HIV seropositive women
Controls: HIV seronegative women

Exposure: Number of sexual partners in past year (0-1, 2+)

Potential confounders/effect modifiers: age (15-29, 30+), marital status (married, divorced/widowed, never married), ever used condom (no, yes), education (none, any)

Many possible models:

- 5 explanatory variables so 32 ($=2^5$) possible models (full dataset has 14 variables, so 16384 possible models)

Want to find model which best represents data:

- include key variables in model i.e. exposure and important confounders and effect modifiers

- including variables unnecessarily may increase SEs e.g. if neither confounder nor risk factor but associated with exposure

One possible strategy

(a) Obtain a crude odds ratio for no. of partners.

Case/control	pa2		
	0-1	2+	Total
Control	507	54	561
	90.37	9.63	100.00
Case	148	30	178
	83.15	16.85	100.00
Total	655	84	739
	88.63	11.37	100.00

. mhodds case pa2

OR	p-value	
1.90	0.008	Partners past year (2+ vs 0-1)

b) Investigate crude associations between potential confounders and the outcome and between potential confounders and exposure (among controls in a case-control study).

Crude odds ratios and p-value & each potential confounder:

OR	p-value	
0.73	0.071	Age (30+ vs 15-29)
2.75	<0.001	Marital (divorced/widowed vs married)
0.82	0.565	Marital (never married vs married)
1.13	0.82	Condom use ever (yes vs no)
2.31	<0.001	Education (any vs none/adult only)

pa2	age30 15-29	30+	Total
0-1	332	323	655
	50.69	49.31	100.00
2+	52	32	84
	61.90	38.10	100.00
Total	384	355	739
	51.96	48.04	100.00

(C) From above analyses identify which variables look like they might be actual confounders (and what effect you think they may have)

(d) Use stratification (Mantel-Haenszel), to obtain odds ratios (95% CI) for partners adjusted for each potential confounder and to identify potential effect modification

. mhodds case pa2, by(age30)

No. partners	Case	Control
<2	82	250
2+	21	31

Age<30

OR=2.07

No. partners	Case	Control
<2	66	257
2+	9	23

Age 30+

OR=1.52

Mantel-Haenzel summary OR =1.85

OR (95% CI) for partners	adjusted for	p-value for interaction
1.90 (1.17-3.09)	crude	-
1.85 (1.14-3.02)	age	0.56
1.88 (1.16-3.05)	marital status	0.25
1.90 (1.17-3.10)	condom use	0.90
1.81 (1.09-3.02)	education	0.013

Condom use does not appear to be an important confounder nor is there any evidence that it is an effect modifier or an independent risk factor – drop at this stage?

Age appears to be a rather minor confounder, no evidence of effect modification. Weak evidence that it is a risk factor in its own right. What to do?

Education appears to be a risk factor and possibly an effect modifier – needs further consideration.

Marital status appears to be a risk factor but not an important confounder or effect modifier – keep or drop?

(e) Repeat above analyses using logistic regression.

. xi logistic case i.pa2 i.age30

```
Logit estimates                   Number of obs   =        739
                                  LR chi2(2)      =       9.16
                                  Prob > chi2     =     0.0102

Log likelihood =  -403.4023

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------------------------------
    partners |   1.848311   .4568771     2.485   0.013

         age |   .7518943   .1314336    -1.631   0.103

------------------------------------------------------------------------------
```

(f) Fit model including no. of partners (and any other necessary variables). Add confounders one by one, starting with strongest. Keep confounder in model if effect of partners changes (or if variable is itself strongly associated with the outcome(?)).

Model 1: pa2

```
. xi: logistic case i.pa2

Logit estimates                   Number of obs =    739
                                  LR chi2(1)    =   6.49
                                  Prob > chi2   = 0.0109
Log likelihood = -404.74114
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Ipa2_1 |      1.9031   .4684     2.614   0.009     1.1748      3.0830
------------------------------------------------------------------------------
```

Model 2: pa2 + ed2

```
. xi: logistic case i.pa2 i.ed2

Logit estimates                   Number of obs =    739
                                  LR chi2(2)    =  26.07
                                  Prob > chi2   = 0.0000
Log likelihood = -394.95047
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Ipa2_1 |      1.7857   .4463     2.320   0.020     1.094       2.914
      Ied2_1 |      2.2551   .4278     4.287   0.000     1.555       3.271
------------------------------------------------------------------------------
```

Model 3: pa2 + ed2 + age30

```
. xi: logistic case i.pa2 i.age30 i.ed2

Logit estimates                   Number of obs =   739
                                  LR chi2(3)    =   26.10
                                  Prob > chi2   =  0.0000
Log likelihood =  -394.9342
------------------------------------------------------------------------------
case  Odds Ratio   SE      z     P>|z|   [95% CI]
------------------------------------------------------------------------------
Ipa2_1    1.7897 .4479   2.326   0.020   1.096 2.923
Iage30_1  1.0353 .1991   0.180   0.857    .710 1.509
Ied2_1    2.2882 .4720   4.013   0.000   1.527 3.428
------------------------------------------------------------------------------
```

Summary:

OR (95% CI) for partners	adjusted for
1.90 (1.17-3.08)	crude [Model 1]
1.79 (1.09-2.91)	education [Model 2]
1.79 (1.10-2.92)	age & education [Model 3]

Final model (which estimates the effect of partners adjusted for confounders) could be Model 2 or Model 3.

This approach is often called "forward fitting" (start with simple model & add terms).

Alternative is "backward fitting" (start with complex model & drop terms).

(g) Finally, check for effect modifiers and include if sufficient statistical evidence and plausible.

Model 2b: pa2 + ed2 + pa2*ed2

```
. xi: logistic case i.pa2*i.ed2

Logit estimates                   Number of obs =    739
                                  LR chi2(2) =     33.35
                                  Prob > chi2 =   0.0409

Log likelihood = -391.30984
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Ipa2_1  |   .4052     .3049    -1.20   0.230     .093    1.770
     Ied2_1  |   1.8856    .3747     3.19   0.001     1.277   2.783
  pa2xed2    |   6.4006    5.159     2.30   0.021     1.319   31.07
------------------------------------------------------------------------------
```

Interaction is statistically significant (LRT p=0.007 & Wald p=0.021) but need to consider plausibility.

If interaction thought to be real then present stratum specific estimates. See below.

Otherwise, accept Model 2 as final model.

(h) Final model includes partners and any (strong) confounders/effect modifiers. Can try adding in other variables to check not confounders after adjusting for education.

(i) Perform LRT on partners in final model to estimate effect of partners after adjusting for confounders.

```
. xi: logistic case i.pa2 i.ed2
. est store A
. xi: logistic case i.ed2
. est store B
. lrtest B A
```

This gives chi-square of 5.15 on 1 df (p=0.023).

Conclusion: women reporting 2+ partners were more likely to be HIV positive than those reporting 0-1 partners, even after allowing for confounding effect of education (adjusted odds ratio=1.79, 95% CI:1.09-2.91, p=0.023).

```
. xi: logistic case i.pa2 i.ed2

Logit estimates                   Number of obs =    739
                                  LR chi2(2) =     26.07
                                  Prob > chi2 =   0.0000

Log likelihood = -394.95047

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Ipa2_1  |   1.785714   .4463095     2.32   0.020     1.094017    2.914286
     Ied2_1  |   2.255102   .4278222     4.29   0.000     1.555124    3.270833
------------------------------------------------------------------------------
```

If interaction between pa2 and ed2 is plausible:

Conclusion:

Effect of partners on HIV status varied with education (LRT p=0.003 for interaction):

In educated women, those reporting 2+ partners were more likely to be HIV positive than those reporting 0-1 partners (odds ratio=2.59, 95% CI:1.47-4.57).

In non-educated women, there was not strong evidence that those reporting 2+ partners were more or less likely to be HIV positive than those reporting 0-1 partners (odds ratio=0.41, 95% CI:0.09-1.77).

Does this pattern make sense? (or can one generate a hypothesis to explain why this should be so?)

```
. xi: logistic case i.pa2*i.ed2

Logit estimates                   Number of obs =    739
                                  LR chi2(2) =     33.35
                                  Prob > chi2 =   0.0409

Log likelihood = -391.30984
------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Ipa2_1 |   0.4052     .3049    -1.20   0.230     .0931     1.7704
    Ied2_1 |   1.8856     .3747     3.19   0.001     1.2774    2.7833
 pa2xed2_1 |   6.4006     5.159     2.30   0.021     1.3191    31.0678
------------------------------------------------------------------------------
```

Summary

1. Begin by producing cross-tabulations and calculating percentages and crude OR's:
   - gives a feel for the data
   - identifies likely confounders/effect modifiers
   - informs the modelling process

2. Possible variables to include in a model
   - exposure (s)
   - *a priori* confounders
   - confounders identified from the analysis
   - important independent risk factors
   - effect modifiers

3. Possible methods for building models:
   - build up from simple models to more complex models.
   - start with complex models and reduce to more simple models.

Comparison of methods

Stratification	Regression
In touch with the data	'black box'
Intuitive	Hidden assumptions
Adjustment is for one exposure	Simultaneous adjustment of all exposures
Ok for <3 confounders	Ok for many confounders
Stratum-specific estimates allow eyeball assessment of interaction	Stratum-specific estimates not given automatically
Interactions between confounders included automatically	Fewer parameters if interactions not specified
Performed by hand	Computer needed
Recommended for initial analysis	Recommended for many exposures

Practical Session

Using the ovarian.dta dataset, follow the modelling strategy in the lecture to investigate whether nullgravidity (having never been pregnant) is a risk factor for ovarian cancer. Consider the confounding and modifying effects of age, social class, smoking and oral contraceptive use.

```
-------------------------------------------------------------------------------
----------
       log:  H:\MSMSC\SME98\ovarca.log
  log type:  text
 opened on:   5 Feb 2004, 11:37:55

. do "C:\Documents and Settings\encdchig.000\Local Settings\Temp\STD00000000.tmp"

.
. tab caco nulligra,row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |       nulligra
      caco | Ever preg  Never pre |     Total
-----------+----------------------+----------
   Control |       306         23 |       329
           |     93.01       6.99 |    100.00
-----------+----------------------+----------
      Case |       167         36 |       203
           |     82.27      17.73 |    100.00
-----------+----------------------+----------
     Total |       473         59 |       532
           |     88.91      11.09 |    100.00


. *** 17.7% of cases never pregnant
. *** 7.0% of controls never pregnant
.
. mhodds caco nulligra

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

    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       2.868003      14.67        0.0001         1.630863   5.043614
    ------------------------------------------------------------------

. *** odds of being nulligravid are 2.87 higher (p<0.001) among cases
than controls
. *

. tab caco agegp, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |                    agegp
      caco |     30-39      40-49      50-59      60-64 |     Total
-----------+--------------------------------------------+----------
   Control |        80         87         90         72 |       329
           |     74.07      69.05      64.29      45.57 |     61.84
-----------+--------------------------------------------+----------
      Case |        28         39         50         86 |       203
           |     25.93      30.95      35.71      54.43 |     38.16
-----------+--------------------------------------------+----------
     Total |       108        126        140        158 |       532
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  27.7037   Pr = 0.000

. mhodds caco agegp

Score test for trend of odds with agegp
```

```
(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in agegp)

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   1.487168      24.29        0.0000         1.270047   1.741407
------------------------------------------------------------------------------

. *** odds of ovarian cancer is higher among older women
. *
. mhodds caco agegp, c(2,1)

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

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   1.280788       0.72        0.3975         0.720916   2.275462
------------------------------------------------------------------------------

. mhodds caco agegp, c(3,1)

Maximum likelihood estimate of the odds ratio
Comparing agegp==3 vs. agegp==1

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   1.587302       2.70        0.1004         0.910149   2.768256
------------------------------------------------------------------------------

. mhodds caco agegp, c(4,1)

Maximum likelihood estimate of the odds ratio
Comparing agegp==4 vs. agegp==1

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   3.412698      21.20        0.0000         1.957364   5.950100
------------------------------------------------------------------------------

.
. tab caco soc, col chi

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

           |               soccl
      caco |      I&II        III         IV&V |     Total
-----------+---------------------------------+----------
   Control |       100        162           67 |       329
           |     53.76      63.53        73.63 |     61.84
-----------+---------------------------------+----------
      Case |        86         93           24 |       203
           |     46.24      36.47        26.37 |     38.16
-----------+---------------------------------+----------
     Total |       186        255           91 |       532
           |    100.00     100.00       100.00 |    100.00

          Pearson chi2(2) =  10.8071   Pr = 0.005


. mhodds caco soc

Score test for trend of odds with soc

(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in soc)
```

```
------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   0.657803      10.79        0.0010         0.512313   0.844611
------------------------------------------------------------------------------

. mhodds caco soc, c(2,1)

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

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   0.667528       4.24        0.0394         0.453245   0.983118
------------------------------------------------------------------------------

. mhodds caco soc, c(3,1)

Maximum likelihood estimate of the odds ratio
Comparing soc==3 vs. soc==1

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   0.416522      10.03        0.0015         0.238061   0.728766
------------------------------------------------------------------------------

. *** odds of ovarian cancer is lower among social classes (III, IV,
and V)
. *

. tab caco ocuse, col chi

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

           |        everoc
      caco |     Never       ever |     Total
-----------+----------------------+----------
   Control |       250         79 |       329
           |     58.96      73.15 |     61.84
-----------+----------------------+----------
      Case |       174         29 |       203
           |     41.04      26.85 |     38.16
-----------+----------------------+----------
     Total |       424        108 |       532
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   7.3404   Pr = 0.007


. mhodds caco ocuse

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

------------------------------------------------------------------------------
 Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
------------------------------------------------------------------------------
   0.527426       7.33        0.0068         0.329255   0.844873
------------------------------------------------------------------------------

. *** odds of ovarian cancer is lower among ever oc users
. *

. tab caco smokgp3, col chi

+--------------------+
| Key                |
|--------------------|
```

```
|     frequency     |
| column percentage |
+-------------------+

           |              smokgp3
      caco |     Never         Ex    Current |     Total
-----------+---------------------------------+----------
   Control |       168         80         81 |       329
           |     61.99      64.00      59.56 |     61.84
-----------+---------------------------------+----------
      Case |       103         45         55 |       203
           |     38.01      36.00      40.44 |     38.16
-----------+---------------------------------+----------
     Total |       271        125        136 |       532
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   0.5497   Pr = 0.760

. mhodds caco smokgp3

Score test for trend of odds with smokgp3

(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in smokgp3)

    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       1.040671       0.14        0.7082         0.844577   1.282295
    ------------------------------------------------------------------

. mhodds caco smokgp3, c(1,0)

Maximum likelihood estimate of the odds ratio
Comparing smokgp3==1 vs. smokgp3==0

    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       0.917476       0.15        0.7015         0.590547   1.425393
    ------------------------------------------------------------------

. mhodds caco smokgp3, c(2,0)

Maximum likelihood estimate of the odds ratio
Comparing smokgp3==2 vs. smokgp3==0

    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       1.107515       0.23        0.6350         0.726336   1.688737
    ------------------------------------------------------------------

. *** odds of ovarian cancer unrelated to smoking status
. *
. tab nulligra agegp if caco==0, row chi

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |                    agegp
  nulligra |     30-39     40-49     50-59     60-64 |     Total
-----------+--------------------------------------------+----------
Ever pregnant |        70         80         85         71 |       306
           |     22.88      26.14      27.78      23.20 |    100.00
-----------+--------------------------------------------+----------
Never pregnant |        10          7          5          1 |        23
           |     43.48      30.43      21.74       4.35 |    100.00
-----------+--------------------------------------------+----------
     Total |        80         87         90         72 |       329
           |     24.32      26.44      27.36      21.88 |    100.00
```

```
          Pearson chi2(3) =   7.6433   Pr = 0.054
```

. * women who have had a pregnancy are generally older**

```
. *
. tab nulligra soc if caco==0, row chi
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

nulligra	soccl I&II	III	IV&V	Total
Ever pregnant	95	151	60	306
	31.05	49.35	19.61	100.00
Never pregnant	5	11	7	23
	21.74	47.83	30.43	100.00
Total	100	162	67	329
	30.40	49.24	20.36	100.00

```
          Pearson chi2(2) =   1.8510   Pr = 0.396
```

. * no strong relationship between nulligravidity and social class**

```
. *
. tab nulligra ocuse if caco==0, row chi
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

nulligra	everoc Never	ever	Total
Ever pregnant	229	77	306
	74.84	25.16	100.00
Never pregnant	21	2	23
	91.30	8.70	100.00
Total	250	79	329
	75.99	24.01	100.00

```
          Pearson chi2(1) =   3.1794   Pr = 0.075
```

. * women who have had a pregnancy are more likely to have used OC's**

```
. *
. tab nulligra smokgp3 if caco==0, row chi
```

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

nulligra	smokgp3 Never	Ex	Current	Total
Ever pregnant	155	74	77	306
	50.65	24.18	25.16	100.00
Never pregnant	13	6	4	23
	56.52	26.09	17.39	100.00
Total	168	80	81	329
	51.06	24.32	24.62	100.00

```
          Pearson chi2(2) =   0.7010   Pr = 0.704
```

. * no strong relationship between nulligravidity and smoking**

```
. *
. mhodds caco nulligra, by(agegp)

Maximum likelihood estimate of the odds ratio
Comparing nulligra==1 vs. nulligra==0
by agegp

------------------------------------------------------------------------------
    agegp | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+-------------------------------------------------------------------
    30-39 |   5.250000          11.67         0.0006         1.81163   15.21416
    40-49 |   1.306122           0.16         0.6857         0.35699    4.77866
    50-59 |   4.794872           8.53         0.0035         1.49816   15.34600
    60-64 |   8.298701           5.41         0.0200         0.98151   70.16557
------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for agegp
    ------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------
       3.965292      22.16        0.0000         2.132977   7.371640
    ------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(3)    =     3.97
                                     Pr>chi2    =   0.2647
```

. * Odd's ratio for nulligravidity adjusted for age is 3.97 (crude was 2.87)**

. * No evidence (p=0.26) that age modifies the effect of nulligravidity**

. * OR for effect of nulligravidty among women 30-39 years is 5.25**

. * OR for effect of nulligravidity among women 40-49 years is 1.31**

. * OR for effect of nulligravidity among women 50-59 years is 4.79**

. * OR for effect of nulligravidity among women 60-65 years is 8.30**

```
. *
. mhodds caco nulligra, by(soc)

Maximum likelihood estimate of the odds ratio
Comparing nulligra==1 vs. nulligra==0
by soc

------------------------------------------------------------------------------
      soc | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+-------------------------------------------------------------------
     I&II |   5.757576          13.17         0.0003         1.97355   16.79698
      III |   2.432681           4.54         0.0330         1.04549    5.66042
     IV&V |   0.779221           0.09         0.7672         0.14882    4.08004
------------------------------------------------------------------------------

    Mantel-Haenszel estimate controlling for soc
    ------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------
       2.883756      13.69        0.0002         1.602499   5.189426
    ------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(2)    =     4.59
                                     Pr>chi2    =   0.1010
```

. * Odd's ratio for nulligravidity adjusted for social class is 2.88 (crude was 2.87)**

. * no statistical evidence (p=0.10) that social class is modifying the effect of nulligravidi**

> ty

```
. *
. mhodds caco nulligra, by(ocuse)

Maximum likelihood estimate of the odds ratio
Comparing nulligra==1 vs. nulligra==0
by ocuse
```

ocuse	Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
Never	2.271825	7.56	0.0060	1.24501	4.14550
ever	10.043478	10.10	0.0015	1.72862	58.35371

Mantel-Haenszel estimate controlling for ocuse

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
2.709792	13.52	0.0002	1.557767	4.713780

```
Test of homogeneity of ORs (approx): chi2(1)  =    2.78
                                     Pr>chi2  =  0.0953
```

. * Odd's ratio for nulligravidity adjusted for OC use is 2.71 (crude was 2.87)**
. * no statistical evidence (p=0.10) that OC use is modifying the effect of nulligravidity**

```
. *
. mhodds caco nulligra, by(smokgp)

Maximum likelihood estimate of the odds ratio
Comparing nulligra==1 vs. nulligra==0
by smokgp3
```

smokgp3	Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
Never	2.873031	8.11	0.0044	1.34320	6.14527
Ex	3.083333	4.23	0.0398	0.99501	9.55459
Current	2.807292	2.65	0.1033	0.76675	10.27826

Mantel-Haenszel estimate controlling for smokgp3

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
2.911029	14.97	0.0001	1.650523	5.134184

```
Test of homogeneity of ORs (approx): chi2(2)  =    0.01
                                     Pr>chi2  =  0.9930
```

. * Odd's ratio for nulligravidity adjusted for smoking is 2.91 (crude was 2.87)**
. * no statistical evidence (p=0.99) that OC use is modifying the effect of nulligravidity**
. *
. *
. * Age has a very strong confounding effect (3.97 v 2.87).**
. * OC use has a modest confounding effect (2.71 v 2.87)**
. * Social class has no confounding effect (2.88 v 2.87)**
. * Smoking has no confounding effect (2.91 v 2.87)**

. *
. *

Check similar results with regression

```
. xi: logistic caco nulligra

Logistic regression                         Number of obs   =        532
                                            LR chi2(1)      =      14.22
                                            Prob > chi2     =     0.0002
Log likelihood = -346.58177                 Pseudo R2       =     0.0201
```

caco	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nulligra	2.868003	.8137856	3.71	0.000	1.644577	5.001555

```
. xi: logistic caco nulligra i.agegp
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(4)      =      48.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -329.19902                       Pseudo R2       =     0.0692

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   3.911892    1.18137     4.52   0.000     2.164356    7.070417
   _Iagegp_2 |   1.571705   .4812589     1.48   0.140      .862445    2.864247
   _Iagegp_3 |   1.886992   .5572543     2.15   0.032     1.057788    3.366212
   _Iagegp_4 |   4.459552   1.289266     5.17   0.000     2.530504    7.859146
------------------------------------------------------------------------------

. xi: logistic caco nulligra i.soc
i.soc             _Isoc_1-3           (naturally coded; _Isoc_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(3)      =      24.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -341.50651                       Pseudo R2       =     0.0344

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |    2.81215   .8068163     3.60   0.000     1.602602    4.934593
     _Isoc_2 |   .6863573   .1366627    -1.89   0.059     .4645847    1.013995
     _Isoc_3 |    .422796   .1198038    -3.04   0.002      .242624    .7367634
------------------------------------------------------------------------------

. xi: logistic caco nulligra i.ocuse
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(2)      =      20.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -343.28307                       Pseudo R2       =     0.0294

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   2.776593   .7924774     3.58   0.000     1.586967    4.857991
   _Iocuse_1 |   .5471745   .1318922    -2.50   0.012     .3411538    .8776097
------------------------------------------------------------------------------

. xi: logistic caco nulligra i.smokgp3
i.smokgp3         _Ismokgp3_0-2       (naturally coded; _Ismokgp3_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(3)      =      15.10
                                                  Prob > chi2     =     0.0017
Log likelihood = -346.13821                       Pseudo R2       =     0.0214

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   2.912463   .8289393     3.76   0.000     1.667221    5.087773
 _Ismokgp3_1 |   .9168343   .2090395    -0.38   0.703     .5864293    1.433395
 _Ismokgp3_2 |   1.162775   .2535033     0.69   0.489     .7584382     1.78267
------------------------------------------------------------------------------

.
. xi:logistic caco i.agegp*nulligra
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)
i.agegp*nulli-a   _IageXnulli_#       (coded as above)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(7)      =      52.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -327.21716                       Pseudo R2       =     0.0748

------------------------------------------------------------------------------
```

```
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
   _Iagegp_2 |   1.914062   .6571039     1.89   0.059      .976643    3.751253
   _Iagegp_3 |   2.007353   .6783292     2.06   0.039     1.035104    3.892814
   _Iagegp_4 |   4.744718   1.529075     4.83   0.000     2.522869    8.923315
    nulligra |        5.25   2.677598     3.25   0.001     1.932105    14.26553
_IageXnull~2 |   .2487852   .2072618    -1.67   0.095     .0486063    1.273376
_IageXnull~3 |   .9133089   .7005942    -0.12   0.906     .2030793    4.107425
_IageXnull~4 |   1.580705   1.869179     0.39   0.699     .1557067    16.04702
------------------------------------------------------------------------------

. est store A

. xi:logistic caco i.agegp nulligra
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(4)      =      48.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -329.19902                       Pseudo R2       =     0.0692

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
   _Iagegp_2 |   1.571705   .4812589     1.48   0.140      .862445    2.864247
   _Iagegp_3 |   1.886992   .5572543     2.15   0.032     1.057788    3.366212
   _Iagegp_4 |   4.459552   1.289266     5.17   0.000     2.530504    7.859146
    nulligra |   3.911892    1.18137     4.52   0.000     2.164356    7.070417
------------------------------------------------------------------------------

. est store B

. lrtest B A

likelihood-ratio test                                 LR chi2(3)  =      3.96
(Assumption: B nested in A)                           Prob > chi2 =    0.2654

.
. xi:logistic caco i.soc*nulligra
i.soc             _Isoc_1-3           (naturally coded; _Isoc_1 omitted)
i.soc*nulligra    _IsocXnulli_#       (coded as above)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(5)      =      29.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -339.04622                       Pseudo R2       =     0.0414

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     _Isoc_2 |   .7530604   .1596747    -1.34   0.181     .4969879    1.141074
     _Isoc_3 |   .5277778   .1563853    -2.16   0.031     .2952789    .9433434
    nulligra |   5.757576   3.023015     3.33   0.001     2.057388     16.1125
_IsocXnull~2 |   .4225183   .2857223    -1.27   0.203     .1122618    1.590227
_IsocXnull~3 |   .1353383   .1340229    -2.02   0.043     .0194309     .942647
------------------------------------------------------------------------------

. est store A

. xi:logistic caco i.soc nulligra
i.soc             _Isoc_1-3           (naturally coded; _Isoc_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(3)      =      24.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -341.50651                       Pseudo R2       =     0.0344

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     _Isoc_2 |   .6863573   .1366627    -1.89   0.059     .4645847    1.013995
     _Isoc_3 |    .422796   .1198038    -3.04   0.002      .242624    .7367634
    nulligra |    2.81215   .8068163     3.60   0.000     1.602602    4.934593
------------------------------------------------------------------------------

. est store B
```

```
. lrtest B A

likelihood-ratio test                                 LR chi2(2)  =      4.92
(Assumption: B nested in A)                           Prob > chi2 =    0.0854

.
. xi:logistic caco i.ocuse*nulligra
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)
i.ocuse*nulli~a   _IocuXnulli_#       (coded as above)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(3)      =      23.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -341.75183                       Pseudo R2       =     0.0338

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iocuse_1 |    .475018   .1236658    -2.86   0.004     .2851734    .7912455
    nulligra |   2.271825   .6900654     2.70   0.007     1.252631    4.120281
_IocuXnull~1 |   4.420885   3.992026     1.65   0.100     .7531484    25.95003
------------------------------------------------------------------------------

. est store A

. xi:logistic caco i.ocuse nulligra
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(2)      =      20.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -343.28307                       Pseudo R2       =     0.0294

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iocuse_1 |   .5471745   .1318922    -2.50   0.012     .3411538    .8776097
    nulligra |   2.776593   .7924774     3.58   0.000     1.586967    4.857991
------------------------------------------------------------------------------

. est store B

. lrtest B A

likelihood-ratio test                                 LR chi2(1)  =      3.06
(Assumption: B nested in A)                           Prob > chi2 =    0.0801

.
. xi:logistic caco i.smokgp3*nulligra
i.smokgp3         _Ismokgp3_0-2       (naturally coded; _Ismokgp3_0 omitted)
i.smok~3*null~a   _IsmoXnulli_#       (coded as above)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(5)      =      15.12
                                                  Prob > chi2     =     0.0099
Log likelihood = -346.13089                       Pseudo R2       =     0.0214

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ismokgp3_1 |   .9084989   .2221481    -0.39   0.695     .5625832    1.467108
 _Ismokgp3_2 |   1.164137     .26628     0.66   0.506     .7435389    1.822655
    nulligra |   2.873031   1.095612     2.77   0.006      1.36063    6.066531
_IsmoXnull~1 |   1.073199   .7314242     0.10   0.917     .2821989    4.081362
_IsmoXnull~2 |   .9771186   .7390663    -0.03   0.976     .2218816    4.303019
------------------------------------------------------------------------------

. est store A

. xi:logistic caco i.smokgp nulligra
i.smokgp3         _Ismokgp3_0-2       (naturally coded; _Ismokgp3_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(3)      =      15.10
                                                  Prob > chi2     =     0.0017
Log likelihood = -346.13821                       Pseudo R2       =     0.0214
```

```
------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ismokgp3_1 |   .9168343   .2090395    -0.38   0.703     .5864293    1.433395
 _Ismokgp3_2 |   1.162775   .2535033     0.69   0.489     .7584382     1.78267
    nulligra |   2.912463   .8289393     3.76   0.000     1.667221    5.087773
------------------------------------------------------------------------------

. est store B

. lrtest B A

likelihood-ratio test                                 LR chi2(2)  =      0.01
(Assumption: B nested in A)                           Prob > chi2 =    0.9927

.
. *** Regression results are similar to Mantel-Haenszel results
.
. *** Include nulligravidity in model and age (as apriori confounder)
. xi: logistic caco nulligra i.agegp
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(4)      =      48.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -329.19902                       Pseudo R2       =     0.0692

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   3.911892    1.18137     4.52   0.000     2.164356    7.070417
   _Iagegp_2 |   1.571705   .4812589     1.48   0.140      .862445    2.864247
   _Iagegp_3 |   1.886992   .5572543     2.15   0.032     1.057788    3.366212
   _Iagegp_4 |   4.459552   1.289266     5.17   0.000     2.530504    7.859146
------------------------------------------------------------------------------

. *** OR for nulligravidity (adjusted for age) is 3.91
. *
. *** Add in OC use as strongest confounder from above
. xi: logistic caco nulligra i.agegp i.ocuse
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(5)      =      49.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -329.16387                       Pseudo R2       =     0.0693

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   3.868687   1.178909     4.44   0.000     2.129001    7.029932
   _Iagegp_2 |   1.552033   .4809408     1.42   0.156     .8455364     2.84885
   _Iagegp_3 |   1.838664   .5715938     1.96   0.050     .9997362    3.381577
   _Iagegp_4 |   4.312728   1.358546     4.64   0.000     2.326036    7.996276
   _Iocuse_1 |   .9308977   .2518276    -0.26   0.791      .547815    1.581867
------------------------------------------------------------------------------

. *** OR for nulligravidity (adjusted for age and OC use) is 3.87
.
. *** Add Social class into model
. xi: logistic caco nulligra i.agegp i.ocuse i.soc
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)
i.soc             _Isoc_1-3           (naturally coded; _Isoc_1 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(7)      =      60.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -323.64196                       Pseudo R2       =     0.0850

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   3.804953   1.177125     4.32   0.000     2.074991     6.97722
```

```
   _Iagegp_2 |   1.615132   .5060764     1.53   0.126     .8739712    2.984824
   _Iagegp_3 |   1.907568    .601916     2.05   0.041     1.027758    3.540536
   _Iagegp_4 |   4.466222   1.431569     4.67   0.000     2.382883    8.371013
   _Iocuse_1 |   .8938502   .2460411    -0.41   0.684     .5211504    1.533086
     _Isoc_2 |   .6102837   .1274258    -2.37   0.018     .4053246    .9188839
     _Isoc_3 |   .4118381   .1205483    -3.03   0.002     .2320455     .730937
------------------------------------------------------------------------------

. *** Odds ratio for nulligravidity is reduced to 3.80
.
. *** Finally Add smoking into final model
. xi: logistic caco nulligra i.agegp i.ocuse i.soc i.smokgp3
i.agegp           _Iagegp_1-4         (naturally coded; _Iagegp_1 omitted)
i.ocuse           _Iocuse_0-1         (naturally coded; _Iocuse_0 omitted)
i.soc             _Isoc_1-3           (naturally coded; _Isoc_1 omitted)
i.smokgp3         _Ismokgp3_0-2       (naturally coded; _Ismokgp3_0 omitted)

Logistic regression                               Number of obs   =        532
                                                  LR chi2(9)      =      61.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -322.84738                       Pseudo R2       =     0.0872

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nulligra |   3.911952   1.218059     4.38   0.000     2.124986    7.201632
   _Iagegp_2 |   1.618791   .5082242     1.53   0.125     .8748902    2.995215
   _Iagegp_3 |   1.889476   .5979376     2.01   0.044     1.016185    3.513257
   _Iagegp_4 |   4.536853   1.461519     4.69   0.000     2.412934    8.530294
   _Iocuse_1 |   .8993467    .248724    -0.38   0.701     .5230199    1.546451
     _Isoc_2 |    .606513   .1268992    -2.39   0.017      .402481    .9139763
     _Isoc_3 |   .4046065   .1188063    -3.08   0.002     .2275574     .719407
 _Ismokgp3_1 |   .8439415    .201293    -0.71   0.477     .5287954    1.346905
 _Ismokgp3_2 |   1.188492   .2722591     0.75   0.451     .7585833     1.86204
------------------------------------------------------------------------------

. *** Odds ratio goes back up to 3.91
.
. *** The only strong confounder is age
.
```

SESSION 10: SOLUTIONS TO EXERCISES IN THE LECTURE

Exercise 1

a). From the raw data it is possible to calculate the odds of infection for each combination of area and age group. For example, for those in **agegrp** 0 who live in the savannah (**area** = 0), the odds of infection are 16/77=0.208. In the table below, fill in the missing odds of infection and odds ratios for area in each age group.

	odds		
agegrp	**area** = 0	**area** = 1	odds ratio
0	0.208	0.380	1.828
1	0.440	1.116	2.536
2	1.447	4.400	3.041
3	2.182	10.32	4.730

Exercise 2

Fill in the empty cells in the table below with the fitted odds from the model:

Odds of infection = Baseline × Area × Age group

	Area		
Agegrp	0	1	**Odds ratio**
0	0.147	0.453	3.08
1	0.382	1.178	3.08
2	1.435	4.421	3.08
3	2.593	7.995	3.08

SESSION 9: SOLUTIONS TO EXERCISES IN LECTURE

Exercise 1: *Fill in the prevalence, odds and log odds of microfilarial infection in the forest and savannah areas.*

	Savannah	Forest	Total
Prevalence	51.3%	71.8	63.1%
Odds	1.052	2.540	1.712
log odds	0.051	0.932	0.538

Exercise 2: *Calculate a 95% confidence interval for the log odds ratio for forest vs savannah and confirm that this corresponds to the STATA output. Convert this to a 95% CI for the odds ratio*

SE (log odds ratio): $= \sqrt{(1/267 + 1/213 + 1/281 + 1/541)} = \sqrt{(0.0138)} = 0.11767$

95% CI for the log odds ratio: $= 0.88 +/- 1.96 \times 0.118 = (0.65, 1.11)$

95% CI for odds ratio: $= (1.92\text{-}3.04)$

Exercise 3: *Calculate the values of L_0 and L_1 and hence the likelihood ratio statistic.*

$L_0 = -857.029$

$L_1 = -379.666 + -448.851 = -828.517$

$L_1 - L_0 = 28.513$ and $LRS = 2(L_1 - L_0) = 2 \times 28.513 = 57.025$

Exercise 4: *Complete this table by calculating the odds, log odds, odds ratios (compared to age group 0) and log odds ratios for age groups 2 and 3.*

	Age group (yrs)			
	5-9	**10-19**	**20-39**	**≥40**
Odds	0.29	0.83	2.392	4.725
log odds	-1.221	-0.184	0.872	1.553
odds ratio	1.00	2.82	8.11	16.02
log OR	0	1.037	2.093	2.774